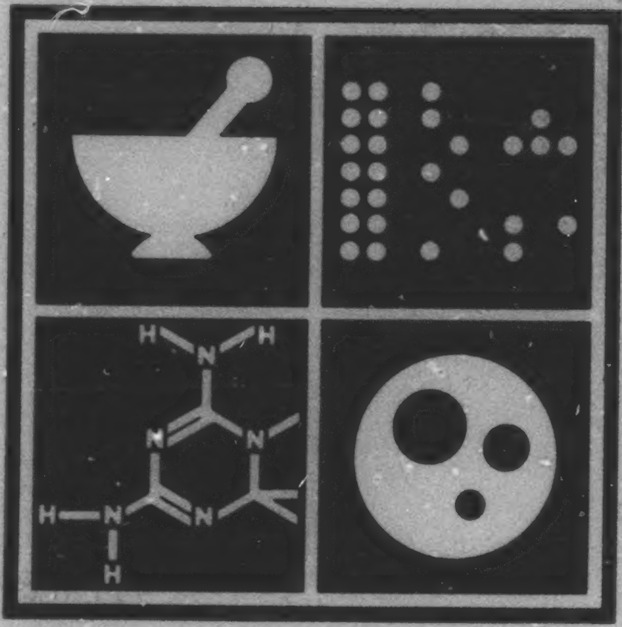
JUNE 86-11391

Bx 86-53



PROPHET

STATISTICS

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service
National Institutes of Health

READEX FICHE best copy available

P 94/3/985

PROPHET STATISTICS

A User's Guide to Statistical Analysis on
the PROPHET System

NEWARK PUBLIC LIBRARY

MAR 14 1986

U. S. GOVT. PUBS.
DEPOSITORY 376

DIVISION OF RESEARCH RESOURCES

NATIONAL INSTITUTES OF HEALTH

Prepared by Bolt Beranek and Newman Inc., Cambridge, Mass.
Under Contract # N01-RR-5-2101
Biomedical Research Technology Program
Division of Research Resources
National Institutes of Health

REVISED SEPTEMBER 1985

HE 20.3008:

PREFACE

PROPHET Statistics is designed to provide a comprehensive and integrated description of the scope and use of the statistics facilities available within the PROPHET System.

PROPHET is a national computer research resource developed specifically for use by biomedical scientists studying chemical and biological systems. PROPHET combines techniques in interactive graphics, data management, and numerical computation into an "automated laboratory notebook". Communicating with PROPHET via higher-level English-like commands, scientists can create, modify, display and manipulate a variety of standard entities such as data tables, graphs and molecular models.

The PROPHET statistics repertoire includes a wide range of statistical methods for analyzing research data, including t-tests, nonparametric tests, analysis of variance, multisample comparisons, frequency analysis and a variety of regression techniques. The design of PROPHET makes these statistics tests readily accessible through:

- direct use of the data stored in PROPHET tables and graphs; (there is no need to convert data to special or unusual formats)
- simple commands that describe the operation; for example, COMPARE, FIT LINE and DISPLAY STATS
- helpful prompts which serve as guides to the analysis appropriate to your data, e.g., "ARE YOU COMPARING 1, 2, OR MORE THAN 2 SAMPLES?", and
- extensive tabular and textual reports of the results of the tests.

Overview of the Manual

The manual will be useful as a teaching guide for those who are just learning PROPHET statistics and will also serve as a reference guide for the experienced PROPHET user. It is not, however, a tutorial either in PROPHET or in the use of statistics.

We assume a general familiarity and background in statistics; we also assume that the reader has knowledge of basic PROPHET facilities, such as making tables and graphs.

Preface

If you have no statistics background, you should first read one of the introductory texts cited in the References at the end of this manual. For an introduction to PROPHET usage, you should read the PROPHET *Primer* and *User's Manual*.

Finally, if you want to read further about statistical analyses, you should consult one of the more advanced texts cited in the References. Zar's *Biostatistical Analysis*, from which the data in some examples were taken, is recommended. Needless to say, consulting with an experienced biostatistican is often the most efficient and rewarding approach to solving your statistics problems.

Our approach in this manual is to provide explanatory material with the descriptions of the statistical tests. To help scientists analyze their data with PROPHET we also provide examples of typical statistical problems to illustrate principles and guide the user in the choice of methodology. In addition, the mathematical and statistical terms are succinctly defined, and the mathematical formulae used in the various tests are included.

Organization

The manual has three major parts.

- Part I provides a summary of the top level command formats for the statistical commands and procedures in PROPHET. This part will be useful as a reference card for those who are proficient in statistics and PROPHET. References to sections in Part II of the manual that contain further discussion of the commands are included in this summary. The Methodology Index in Part I provides an alphabetical listing of the statistical tests, the corresponding PROPHET commands and references to the pertinent sections in Part II.
- Part II explains in detail all of the statistics tests in PROPHET, along with examples and descriptions of the statistical methodology.
- Part III is designed for those who wish to write their own PROPHET statistical procedures. This part lists the "quiet entry" calls to the various statistical procedures in alphabetical order, with descriptions of input arguments, type of output, and so on.

Finally, the manual provides references to appropriate statistical texts where more extensive discussion of the tests may be found, in particular those texts whose format is consistent with the PROPHET approach. A glossary of symbols used is also included.

CONTENTS

Page

PREFACE

COMMAND SUMMARY

METHODOLOGY INDEX

INTRODUCTION

1. DESCRIPTIVE STATISTICS
   - 1.1 Introduction 1-1
   - 1.2 Basic Measures 1-2
   - 1.3 PROPHET Usage: STATS 1-4
   - 1.4 PROPHET Usage: ADD STATS and DELETE STATS 1-9
   - 1.5 Additional Measures 1-11
   - 1.6 PROPHET Usage: MEASURES 1-14
   - 1.7 PROPHET Usage: SELECT STATS 1-15
   - 1.8 PROPHET Usage: LOGSTATS and LOGMEASURES 1-17
   - 1.9 PROPHET Usage: Individual Measures 1-18
   - 1.10 PROPHET Usage: Graphs 1-19

2. DISTRIBUTION TESTING
   - 2.1 Introduction 2-1
   - 2.2 PROPHET Usage: Making Histograms 2-2
   - 2.3 Normal Distribution, Skewness and Kurtosis 2-5
   - 2.4 PROPHET Usage: TEST NORMALITY and TEST NORMALSHAPE .. 2-9
   - 2.5 Testing Other Distributions 2-11
   - 2.6 PROPHET Usage: TEST DISTRIBUTION 2-13

3. ONE- AND TWO-SAMPLE HYPOTHESIS TESTING
   - 3.1 Statistical Hypothesis Testing 3-1
   - 3.2 Parametric and Nonparametric Methods 3-2
   - 3.3 PROPHET Usage: COMPARE 3-3
   - 3.4 Parametric Tests for One and Two Samples 3-6
   - 3.5 Nonparametric Tests 3-13
   - 3.6 PROPHET Usage: TEST VARIANCES 3-16

4. MULTISAMPLE HYPOTHESIS TESTING

4.1 Introduction 4-1
4.2 PROPHET Usage: COMPARE 4-2
4.3 Principles of Analysis of Variance 4-6
4.4 Assumptions 4-9
4.5 One-way Analysis of Variance 4-10
4.6 Two-way Analysis of Variance 4-13
4.7 Multiple Comparisons for Parametric Tests 4-16
4.8 Kruskal-Wallis Test 4-19
4.9 Friedman's Test 4-21
4.10 Nonparametric Multiple Comparison Tests 4-23
4.11 PROPHET Usage: TEST VARIANCES 4-25

5. FREQUENCY ANALYSIS

5.1 Contingency Tables 5-1
5.2 PROPHET Usage: CHISQ - χ^2 Test for Independence 5-2
5.3 CHISQ Statistics 5-3
5.4 PROPHET Usage: FISHER 5-4
5.5 PROPHET Usage: CROSSTAB and STATPAK 5-9
5.6 STATPAK Statistics 5-14
5.7 PROPHET Usage: CHISQ - χ^2 Test for Goodness of Fit ... 5-20

6. REGRESSION ANALYSIS

6.1 Introduction 6-1
6.2 Principles and Assumptions of the Regression Technique 6-4
6.3 PROPHET Usage: Introduction 6-5
6.4 Simple Linear Regression 6-6
6.5 PROPHET Usage: FIT LINE and CONFCURVS 6-13
6.6 Multiple Linear Regression 6-16
6.7 PROPHET Usage: MULREG 6-18
6.8 Polynomial Regression 6-22
6.9 PROPHET Usage: FIT POLYNOMIAL 6-23
6.10 PROPHET Usage: FIT FUNCTION 6-26
6.11 PROPHET Usage: FITFUN 6-30
6.12 PROPHET Usage: EXPFIT 6-35
6.13 PROPHET Usage: HYPERBOLIC 6-40

7. MULTIFACTOR ANAYLSIS OF VARIANCE

7.1 Introduction 7-1
7.2 PROPHET Usage: ANVAR 7-3
7.3 Perparing the Data for ANVAR Using ARAFIL 7-10
7.4 PROPHET Usage: ARAFIL 7-12

8. ANALYSIS OF COVARIANCE

8.1 Introduction 8-1
8.2 Basic Models in Analysis of Covariance 8-2
8.3 Computations in Analysis of Covariance 8-3
8.4 Hypothesis Testing in Analysis of Covariance 8-5
8.5 PROPHET Usage: TEST LINES 8-9

9. MISCELLANEOUS TESTS
  9.1 Multivariate Analysis 9-1
  9.2 Hotelling's T^2 9-2
  9.3 PROPHET Usage: HOTELLING 9-4

FUNCTIONS AND PROCEDURES

REFERENCES

GLOSSARY OF SYMBOLS

INDEX

COMMAND SUMMARY

STATS and MEASURES (Section 1)

```
DISPLAY
EAD                      [LOG]STATS    OF |COL[UMNS] OF|  table [FOR-clause]
PRINT                    [LOG]MEASURES    |ROW[S]      |  float array array
PRINTOUT                                  |float array
SET table-name TO


ADD [LOG]STATS      TO |COL[UMNS] OF|  table [FOR-clause]
    [LOG]MEASURES      |ROW[S]      |
```

```
        e.g.,  EAD STATS OF COLS OF DATATAB
               SCAN STATS OF DATATAB
               DISPLAY MEASURES OF COLS OF DATATAB
               PRINT LOGSTATS OF DATATAB FOR COLS <>1
               ADD STATS TO ROWS OF DATATAB.
```

STATS computes N, SUM, MEAN, VARIANCE, STDEV, CI95

MEASURES computes N, SUM, MEAN, VARIANCE, STDEV, CI95, CI99, SEM, CV%, MEDIAN, GEOMEAN, MIDRANGE, MINIMUM, MAXIMUM, RANGE, SKEWNESS, KURTOSIS

<u>Note</u>: FOR-clause selects the data from the specified table to compute the statistics on.

SELECT STATS [TO STANDARD] (Section 1.7)

To select measures to be computed by STATS and specify the number of digits to round results to.

```
DELETE STATS    FROM  |COLUMNS  OF|  table
                      |ROWS       |
```

(Section 1.4)

Deletes the measures added to the table by the ADD STATS or ADD MEASURES command.

Command Summary

```
TYPE    MEAN      OF  COLUMN col-id  OF  table              (Section 1.9)
        SUM           ROW    row-id      float array array
        VARIANCE
        STDEV         float array
        MEDIAN
        MINIMUM
        MAXIMUM
```

Computes a single measure of a row, column or all of a table or array.

```
                                                          (Section 2.2)
MAKE TABLE table-name AS DISTRIBUTION OF COLUMN col-id OF table
                                         float array

 [FOR INTERVALS OF  float  STARTING AT  float]
 [                         ENDING            ]
 [                  COLUMN  col-id  OF  table]
 [                  float array              ]

        e.g., MAKE TABLE DISTN AS DISTRIBUTION OF COL 2 OF DATATAB
                     FOR INTERVALS OF 2.STARTING AT 60.
```

Makes distribution table

```
TEST NORMALITY                                            (Section 2.4)
```

Tests whether sample data follow normal distribution.

```
TEST NORMALSHAPE                                          (Section 2.4)
```

Tests skewness and/or kurtosis to determine how the data depart from normality.

```
TEST DISTRIBUTION                                         (Section 2.6)
```

Tests whether sample data follow specified distribution (normal, binomial, Poisson, uniform, chi-square, F and t).

COMPARE (Sections 3 and 4)

- one- and two-sample t-tests.
- completely randomized one-way analysis of variance and multiple comparisons.
- randomized block two-way analysis of variance and multiple comparisons.
- one- and two-sample (paired) sign rank test.
- two-sample rank sum test.
- Kruskal-Wallis (unblocked) test and multiple comparisons.
- Friedman (blocked) test and multiple comparisons.

TEST VARIANCES (Sections 3.6 and 4.11)

- F-test for equality of two variances.
- Levene's or Bartlett's test for homogeneity of variance.

TEST DISPERSIONS (Section 3)

Ansari-Bradley test for equality of two dispersions.

CALL CHISQ(table) (Public Procedure) (Sections 5.2 and 5.7)

Computes χ^2 , coefficient of contingency and Cramer's V from frequency table or computes χ^2 from columns of observed and expected frequencies.

```
CALL FISHER(contingency-table)      (Public Procedure)     (Section 5.4)
```

Fisher's exact test for independence of 2x2 contingency table.

```
CALL CROSSTAB(data-table)           (Public Procedure)     (Section 5.5)
```

Builds contigency table on two variables with option for additional variable control; calls STATPAK automatically.

```
CALL STATPAK(contingency-table)     (Public Procedure)     (Section 5.5)
```

Computes X^2, coefficient of contingency, Cramer's V, lambda, Kendall's tau, Spearman's rho, and gamma on a frequency table.

(Section 6)

```
                       LINE
[fit-text = ]   FIT    POLYNOMIAL              TO
                       FUNCTION [fun-text]

        [COLUMN  x-col  VS  COLUMN  y-col  OF]  table

       |WITH  WEIGHTS   IN COLUMN  wt-col|
       |      VARIANCES                  |


[fit-text = ]   FIT    LINE                   TO [CURVE  curve-number  OF]
                       POLYNOMIAL
                       FUNCTION [fun-text]

            graph       |WITH WEIGHTS   IN COLUMN  wt-col|
                        |     VARIANCES                  |

       e.g., FIT LINE TO DATATAB
             FIT LINE TO COL 4 VS COL 2 OF DATATAB
             FIT POLYNOMIAL TO CURVE 2 OF DATAGRAPH
             FIT FUNCTION TO DATATAB WITH WEIGHTS IN COL 2
```

CALL CONFCURVS (Public Procedure) (Section 6.5)

Confidence intervals for linear regression.

CALL MULREG (Public Procedure) (Section 6.7)

Stepwise multiple regression.

CALL FITFUN (Public Procedure) (Section 6.11)

General functional regression (allows setting up parameter and function tables).

CALL FITFUNSAV

Saves resulting parameter and analysis of variance tables.

CALL FITFUNQK

Manual entry of functional form for single fit.

CALL FITFUNQKSAV

Saves resulting parameter and analysis of variance tables; manual entry.

CALL EXPFIT(data-table) (Public Procedure) (Section 6.12)

Performs peeling and regression analysis with FIT FUNCTION.

CALL HYPERBOLIC(data-table) (Public Procedure) (Section 6.13)

Hyperbolic regression; Michaelis-Menton equation.

CALL %RESIDUALANALYSIS(res-tab) (Section 6)

Graphs residuals from regression analysis. The table res-tab contains X-data in column 1, Y-data in column 2, Y-fitted in column 3 and residuals in column 4. If the table contains more than 4 columns you are asked for the location of the relevant information.

CALL ARAFIL (Public Procedure) (Section 7)

Builds a design table and fills a data array (for use with ANVAR).

CALL ANVAR(data,dimension,anova) (Public Procedure) (Section 7)

2- to 6-way factorial analysis of variance.

data - array from ARAFIL (FLOAT array)
dimension - dimension of data (FIXED)
anova - name of analysis of variance table (TEXT)

TEST LINES (Section 8.5)

Linear covariance analysis.

CALL HOTELLING (Public Procedure) (Section 9)

Computes Hotelling's T^2 to compare two multivariate normal samples.

Mathematical Functions (see User's Manual)

A**B	Exponentiation
A*B	Multiplication
A/B	Division
A+B	Addition
A-B	Subtraction
-A	Unary minus
ABS(A)	Absolute value
ASIN(A)	Arc sine
ACOS(A)	Arc cosine
ATAN(A)	Arc tangent
COS(A)	Cosine (radians)
COSD(A)	Cosine (degrees)
EXP(A)	e raised to the power A
LOG(A)	Natural logarithm
LOG10(A)	Common logarithm
SIN(A)	Sine (radians)
SIND(A)	Sine (degrees)
SQRT(A)	Square root

Methodology Index

Test	PROPHET Usage	Section
ANSARI-BRADLEY Test for Equality of two Dispersions	TEST DISPERSIONS	3
BARTLETT's Test for Homogeneity of Variance	TEST VARIANCES	4
CHI-SQUARED Test	CHISQ	5
Confidence Curves for Linear Regression	CONFCURVS	6
Covariance Analysis	TEST LINES	8
CRAMER'S V	STATPAK	5
D'AGOSTINO'S Test for Normality	TEST NORMALITY (COMPARE)	2 3,4
DUNNETT'S Multiple Comparison Procedure	COMPARE	4
Exponential Regression	EXPFIT	6
F-test for Equality of Two Variances	TEST VARIANCES (COMPARE)	3 3
Factorial Analysis of Variance	ANVAR	7
FISHER'S Exact Test	FISHER	5
FRIEDMAN'S Two-way, Nonparametric Multiple Comparison Test	COMPARE	4
General Functional Regression	FIT FUNCTION FITFUN	6 6
HOTELLING'S T^2	HOTELLING	9
Hyperbolic Regression	HYPERBOLIC	6
KENDALL'S tau	STATPAK	5
KOLMOGOROV-SMIRNOV Distribution Test	TEST DISTRIBUTION	2

Methodology Index

Test	PROPHET Usage	Section
KRUSKAL-WALLIS Nonparametric Multiple Comparison Test	COMPARE	4
Lambda	STATPAK	5
LEVENE'S Test for Homogeneity of Variance	TEST VARIANCES (COMPARE)	4 4
Multiple Linear Regression	MULREG	6
NEWMAN-KEULS Multiple Comparison Test	COMPARE	4
One-way Analysis of Variance	COMPARE	4
Polynomial Regression	FIT POLYNOMIAL	6
Simple Linear Regression	FIT LINE	6
SPEARMAN'S rho	STATPAK	5
t-test, Equal Variances	COMPARE	3
t-test, Unequal Variances	COMPARE	3
t-test, One-sample	COMPARE	3
t-test, Paired	COMPARE	3
Two-way Analysis of Variance	COMPARE	4
WILCOXON'S One-sample Test	COMPARE	3
WILCOXON'S Sign Rank Test	COMPARE	3
WILCOXON-MANN-WHITNEY Test	COMPARE	3
WILK-SHAPIRO Test for Normality	TEST NORMALITY (COMPARE)	2 3

REFERENCES

GENERAL STATISTICS:

Armitage, P. (1971). Statistical Methods in Medical Research. Blackwell Scientific Publications, Oxford.

Bliss, C.J. (1970). Statistics In Biology, 2 Vols. McGraw-Hill, New York.

Brownlee, K.A. (1965). Statistical Theory and Methodology In Science and Engineering. Wiley, New York.

Dixon, W.J. and Massey, F. J. (1969). Introduction to Statistical Analysis. McGraw-Hill, New York.

Snedecor, G.W. and Cochran, W. G. (1967). Statistical Methods. Iowa State University Press, Ames.

Sokal, R.R. and Rohlf, F. J. (1972). Biometry. Freeman, San Francisco.

Swinscow, T.D.V. (1978). Statistics at Square One. Mendip Press, Bath, England.

Zar, J.H. (1974). Biostatistical Analysis. Prentice Hall, Englewood Cliffs, N.J.

REGRESSION ANALYSIS:

Daniel, C. and Wood, F.S. (1971). Fitting Equations To Data. Wiley-Interscience, New York.

Dunn, O.J. and Clark, V.A. (1974). Applied Statistics: Analysis of Variance and Regression. Wiley, New York.

Draper, N.F. and Smith, H. (1966). Applied Regression Analysis. Wiley, New York.

References

ANALYSIS OF VARIANCE:

Cochran, W.G. and Cox, G.M. (1957). Experimental Designs. Wiley, New York.

Dunn, O.J. and Clark, V.A. (1974). Applied Statistics: Analysis of Variance and Regression. Wiley, New York.

Scheffé, H. (1959). The Analysis of Variance. Wiley, New York.

Winer, B.J. (1971). Statistical Principles in Experimental Design. McGraw-Hill, New York.

NONPARAMETRIC STATISTICS:

Conover, W.J. (1971). Practical Nonparametric Statistics. Wiley, New York.

Daniel, W.W. (1978). Applied Nonparametric Statistics. Houghton-Mifflin, Boston.

Hollander, M. and Wolfe, D.A. (1973). Nonparametric Statistical Methods. Wiley, New York.

MULTIVARIATE ANALYSIS:

Anderson, T.W. (1959). An Introduction to Multivariate Statistical Analysis. McGraw-Hill, New York.

Statistics Notation

Univariate Statistics

X_i - observation i of variable X, i=1,...,n

n - number of observations in the sample

$\bar{X}$ - sample mean

s^2 - sample variance

s - sample standard deviation

$s_{\bar{x}}$ - sample standard error of the mean

R_i - rank sum of sample i

CI95 - 95% confidence interval

CI99 - 99% confidence interval

CV% - coefficient of variation expressed as a percent

Regression

X - measured independent variable

X_i - independent variable for the ith observation, i=1,...,n

$\bar{X}$ - mean of the independent variable

X_{ij} - ith observationof the jth independent variable in multiple regression

$\bar{X}_j$ - mean of the jth independent variable

Y - measured dependent variable

Y_i - dependent variable for the ith observation

Statistics Notation

$\bar{Y}$ - mean of the dependent variable

$\hat{Y}_i$ - fitted dependent variable for the ith observation

w_i - weight for observation i

r_{ij} - sample correlation coefficients for variables i and j

n - number of observations

p - number of independent variables in multiple regression

DF - degrees of freedom

R^2 - multiple correlation coefficient squared

SS - sum of squares

MS - mean square

Regression SS - sum of squares due to regression

Residual SS - sum of squares due to error

$s_{y.x}$ - standard deviation of regression

SE(a) - standard error of the intercept

SE(b) - standard error of the slope

Functional forms of various curve fits are as follows:

simple linear fit $Y = \alpha + \beta X$

polynomial fit $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$

exponeential fit $Y = \beta_0 \quad \beta_1 e^{\beta_2 X} + \ldots + \beta_{p-1} e^{\beta_p X}$

multiple regression $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

Statistics Notation

Analysis of Variance

X_{ij} - measured independent variable

$\bar{X}_{i}.,\bar{X}.j,\bar{X}..$ - means in two-way analysis

n_i - number of observations in treatment i

N - total number of observations

$s^2_{\;i}$ - sample variance of treatment i

s^2 - error or residual mean square

(1),a,b,ab - denote treatment combinations in 2*n* factorial designs

A,B,AB - denote main effects of factor A, B, and A x B interaction

Analysis of Covariance

X_{ij} - *jth* measurement of independent variable in regression i

Y_{ij} - *jth* measurement of dependent variable in regression i

SS_i - residual sum of squares for regression i

SS_c - residual sum of squares for common regression

$(s^2_{\;y.x})_p$ - pooled residual mean square

$(s^2_{\;y.x})_c$ - common residual mean square

INTRODUCTION

This introduction briefly discusses some of the most common types of data and methods of analysis that biological investigators use. At the same time, it provides an overview of the PROPHET tests described in Sections 1-9 of Part II.

Types of Data

Prior to doing any statistical analysis, the investigator should thoroughly understand the type of variable under consideration. There are three basic types of variables in biological research; these are measurable variables, ranked variables, and attributes.

Measurable Variables:

Measurable variables can be on either a ratio scale or an interval scale. For both ratio scale data and interval scale data there is a constant size unit between each interval of the measurement scale. For example, the distance between 7 and 9 inches on a ruler is the same as that between 10 and 12 inches. In the case of ratio scale data it is also required that a fixed zero reference be known. This enables such statements as "4 inches is twice as long as 2 inches" to have meaning. An example of an interval scale variable where such a comparison is not valid would be temperature.

Measurable variables, whether on a ratio or interval scale, can be either continuous or discrete. Continuous variables are those whose units of measurement can be made, theoretically at least, infinitely small. Examples of continuous variables are height, weight, voltage, pH, etc. Discrete variables, on the other hand, can by nature assume only certain values. The number of offspring in a litter might be 2, 3, or 4, but not 2.5; the number of successful performances of a task could be 0, 1, or 2, but not 2.25.

Ranked Variables:

Other variables are not easily assigned a measurable, quantitative value, although they may be ranked in a certain logical way. These are called ordinal scale data, or ranked data. For example, the performance of rats in a time trial may be ranked according to the order of their finish, or ranks corresponding to the test grades A, B, etc. may be assigned to a group of students. This ordering provides useful information, although it is clearly erroneous to assume that the rat finishing first performed four times better than the one finishing fourth, or that a student receiving a grade of A performed twice as well on the test as one who received a B.

Attributes:

For the third type of variable, attribute or nominal scale data, it is meaningless to attach a quantitative measure, but different characteristics are distinguished by descriptive information. Examples of attribute data are color, sex, alive or dead, etc.

Types of Statistical Analysis

Initial Analysis:

Confronted with a table of data values, your first objective is usually to obtain an overall picture or summary of the data. When the data are measurable, quantitative statistics such as the mean, variance, etc. are useful. This subject, along with the STATS and the MEASURES expressions, are discussed in Section 1. These measures are often useful with ranked data as well.

All three types of data may be viewed with the use of the histogram. For measurable variables, the data are broken into intervals and the number of observations within each interval are plotted. For nominal scale data, the number of observations which share a given attribute are plotted. Histograms are introduced in Section 1, and the more complex subject of frequency distributions is discussed in Section 2, along with the commands TEST DISTRIBUTIONS, TEST NORMALITY, and TEST NORMALSHAPE.

Hypothesis Testing:

With the basic measures of a sample computed, it is general practice to pursue a course for drawing conclusions or decisions about the data under study. The method is to assume a hypothesis about the population -- for example, that the population mean is equal to a certain value -- and then use the sample in an appropriate test of the hypothesis.

The statistical test is performed to determine if it is appropriate to reject the "null hypothesis", denoted H_0 , specifying equality. Before making the test, strict specifications are set for the conditions that must be met if the null hypothesis is to be considered acceptable.

Of course, the only way that one can ever be certain about the truth or falsity of a statistical hypothesis is to examine the entire population, and, in practice, this can rarely be done. Thus, there is always a possibility that the conclusions about the population reached by certain statistical tests are incorrect.

There are two types of such erroneous conclusions. The first, known as Type I error, is rejecting the null hypothesis when it in fact is true. Its probability, denoted by α, is known as the significance level of the test. When you choose the significance level, you will clearly want it to be small; conventional values for α are 0.05 (i.e., 5%) and 0.01 (i.e., 1%). (Note that a significance level of 0.01 is considered to be higher than 0.05.)

The second type of error is that of accepting the null hypothesis when in fact it is not true. This is known as a Type II error, and its probability is represented by β. In a statistical test the α (significance level) may be selected, but β is typically unknown. But it is known that decreasing α for a fixed sample size (n) results in an increase in β. For α and β both to be decreased, n must be increased.

For a given set of data, some tests are more powerful than others. That is, they have a lower α at a given β. One therefore wants to use the most powerful test appropriate to the data. For example, when samples have a normal distribution, parametric tests give better results than nonparametric tests because, given a significance level, we can expect lesser probability of erroneously accepting the null hypothesis. However, when the data are not normally distributed, a parametric test may have a higher risk of accepting an untrue null hypothesis. In effect, the significance level is much smaller than you set it to be. Testing the distribution of the data (Section 2) is therefore an important step in determining the best test to apply.

If the significance level has been selected as 0.05, and the statistical test indicates rejection of the null hypothesis, you may conclude that either the null hypothesis is not true, or that you have observed a situation that occurs less than 5% of the time by chance alone if the null hypothesis is a true statement about the population. Similarly, the confidence level of the conclusion is $1-\beta$, or 95%. The use of these concepts will be illustrated throughout Part II.

PROBABILITIES OF TYPE I AND TYPE II ERRORS.

	Accept H_0		**Reject H_0**	
	Type of Error	**Probability**	**Type of Error**	**Probability**
H_0 True	No Error	$1-\alpha$	Type I	α
H_0 False	Type II	β	No Error	$1-\beta$

α = probability of Type I error (significance level)

β = probability of Type II error

$1-\beta$ = power of the test

Comparing Samples:

In many experimental situations the objective is to compare two or more samples of data obtained under different experimental conditions. Section 3 introduces hypothesis testing, and illustrates the use of the COMPARE command to perform one- and two-sample tests on both measurable and ranked data. Extensions of these methods in COMPARE to more than two samples (analysis of variance) are described in Section 4.

Independence and Association:

Test procedures for independence and association of two variables on the nominal or ordinal scales (i.e., ranked data or attributes) are discussed in Section 5, which describes the public procedures CHISQ, FISHER, STATPAK and CROSSTAB.

Regression:

In many situations pairs of measurable variables are related, and describing the relationship is of interest. The techniques of regression analysis and the use of the various FIT commands to invoke regression procedures in PROPHET are discussed in Section 6.

Factorial Analysis:

In more complex experimental designs, more than one factor may be influencing an observed effect. Such designs are studied in a factorial analysis of variance; its use in PROPHET with the public procedure ANVAR is presented in Section 7.

Covariance:

Section 8 introduces the simple analysis of covariance, which is useful in studies involving the comparison of simple linear regressions, or in situations where it is desirable to eliminate the effect of a covariate (common variable).

Multivariate Samples:

Finally, in Section 9 Hotelling's T^2 for comparing two multivariate samples is discussed. Multivariate data consist of observations on more than one variable of measurable data.

1. Measures

1. Measures

Section 1

DESCRIPTIVE STATISTICS

1.1 Introduction

The general goal of statistical analysis is to infer conclusions about a population. A population in statistical terms is the collection of all entities that share a set of common attributes, such as age or height. In most cases it is not possible to observe a large or changing population exhaustively. Thus, in practice a subset of the population, known as a sample, is usually studied. A basic assumption is that conclusions about the population can legitimately be inferred from the analysis of the sample.

Usually the first stage of a statistical analysis is to investigate some basic measures of the sample that summarize its properties. The most important are: measures of central tendency or averages, which indicate where the bulk of the data are located; and measures of dispersion, which describe the spread or variation of the individual values. In comparing several samples, these measures may reveal large discrepencies or unusual results that may point out where further investigation is warranted.

1.2 Basic Measures

A. Arithmetic mean

The most commonly used measure of central tendency is the arithmetic mean. This is computed by summing the values of all observations and dividing the result by the total number of observations. In statistical notation, each observation is denoted by X_i, the number of observations by n, and thus the mean, $\overline{X}$, is calculated as

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

The mean of the sample is a "good" estimate of the true mean (μ) of the population from which the sample was drawn. That is, the mean is the most efficient, unbiased and consistent estimate of μ (Zar, pp. 17, 18).

B. Variance

The mean is the average value of a sample; it provides no information about the variability or spread of the data. This is provided by the variance of the sample, s^2, which is an estimate of the average squared distance about the mean, and is calculated as

$$s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$$

Note that s^2 is an unbiased estimator of σ^2, the population variance. That is, if an indefinitely large number of samples of size n were drawn from a single population, in the long run the mean of the s^2 so obtained would equal the true population variance, σ^2. Note that if n were used instead of n-1 in the calculation of s^2, the estimate on average would under-estimate σ^2, particularly when n is small. The term n-1 is known as the degrees of freedom.

C. Standard deviation

One disadvantage of using the variance as a measure of sample dispersion is that variance is expressed as the squared value of the original units of measurement. For example, if the X's are measured in grams then the variance would be expressed in square grams, which has no physical meaning. The standard deviation (s) overcomes this scaling difficulty; it is the positive square root of the variance. Thus,

$$s = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)}$$

D. Confidence intervals

Confidence intervals provide a measure of the precision of the sample estimate of the mean. The most commonly used are the 95% and 99% confidence intervals.

These intervals have the property that if repeated random samples of size n are taken and 95% confidence intervals are computed, then 95% of these intervals will contain the true population mean. Put another way, there is a 95% chance that the calculated confidence interval will enclose the true mean. (Note that this is not the same as saying that there is a 95% probability the population mean lies within the confidence intervals, for the mean is fixed.) The half interval for the 95% confidence level, abbreviated CI95, is

$$\mathrm{CI95} = t_{0.05,\,n-1}\sqrt{s^2/n}$$

where $t_{0.05,n-1}$ is the 0.05 percentage point of the Student's t-distribution. To obtain the lower confidence bound use $\overline{X}$-CI95 and for the upper bound use $\overline{X}$+CI95.

The size of the confidence interval depends on:

- the confidence level chosen
- the magnitude of the variance (or standard deviation), and
- the size of the sample.

The confidence interval will be larger with larger confidence levels, larger variances and smaller sample sizes.

1.3 PROPHET Usage: STATS

Generally, each sample of a data set is stored in either a column or row of a PROPHET table. Most often the data are organized by columns since the data within each column must all be of the same type. But since the organization of your data depends on what analysis is to be done, you can obtain the measures described above on either the rows or columns, or on both, by using the STATS expression. STATS will compute for each column and/or row

- the number of observations, -N-
- the sum, -SUM-
- the mean, -MEAN-
- the variance, -VARIANCE-
- the standard deviation, -STDEV-
- the 95% confidence interval, -CI95-.

STATS on Columns

For example, Figure 1-1 contains the level of pH found in 8 samples of water from each of 4 ponds. One of the measurements was lost from pond 3 (Zar, p. 141).

PH: PH LEVELS OF 4 PONDS 8R X 4C

0	1 POND 1	2 POND 2	3 POND 3	4 POND 4
1.	7.68	7.69	7.74	7.71
2.	7.7	7.7	7.75	7.71
3.	7.72	7.71	7.77	7.74
4.	7.73	7.73	7.78	7.79
5.	7.73	7.74	7.8	7.84
6.	7.76	7.74	7.81	7.85
7.	7.78	7.78	7.81	7.87
8.	7.8	7.81		7.91

Figure 1-1

Since the pH level for each pond is in a column of the table STATS computed on columns is of interest. (STATS can also be computed on rows if your data are organized by rows). The following command would display the measures of the columns of pH as shown in Figure 1-2.

DISPLAY STATS OF COLS OF PH

Each row of Figure 1-2 contains a single measure for each column of pH data.

STATS OF PH: PH LEVELS OF 4 PONDS 6R X 4C

0 STATS	1 POND 1	2 POND 2	3 POND 3	4 POND 4
1. -N-	8.	8.	7.	8.
2. -SUM-	61.9	61.9	54.46	62.42
3. -MEAN-	7.7375	7.7375	7.78	7.8025
4. -VARIANCE-	.001619611	.001649039	.0008010864	.005851201
5. -STDEV-	.0402444	.04060836	.02830347	.07649314
6. -CI95-	.03357936	.03388305	.02620938	.06382481

Figure 1-2

Any of the PROPHET display commands can be used with the STATS expression. For example, EAD, SCAN, PRINT, or PRINTOUT could be used in place of DISPLAY.

STATS on
Rows

The data table WTLOSS (Figure 1-3) records the body weight of several patients treated with a drug which induces weight loss. At several time intervals, in weeks, the patients are weighed.

WTLOSS: WEIGHT OF MALE PATIENTS 7R X 5C

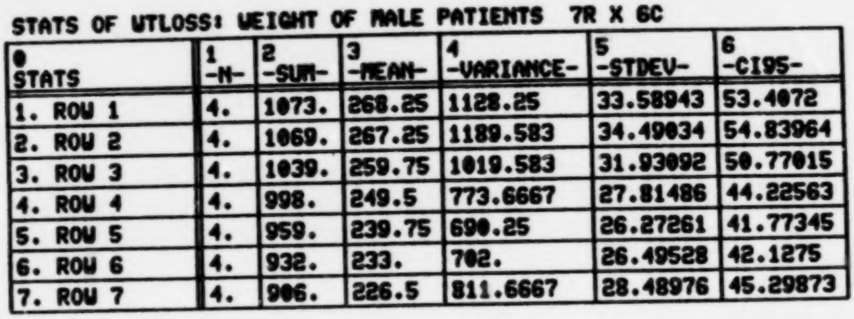
	1 TIME (WKS)	2 PATIENT 1 (LBS)	3 PATIENT 2 (LBS)	4 PATENT 3 (LBS)	5 PATIENT 4 (LBS)
1	0.	253.	232.	310.	278.
2	2.	251.	231.	311.	276.
3	4.	243.	227.	300.	269.
4	8.	232.	221.	281.	264.
5	14.	222.	213.	267.	257.
6	20.	215.	206.	260.	251.
7	28.	210.	196.	258.	242.

Figure 1-3

The table is organized with TIME in column 1 and each patient's weight is recorded in columns 2 to 5. To show the average change in weight vs. time STATS is computed by rows. However, since TIME itself is not a response, it should not be included in the calculations. It is easy to exclude the data in column 1 from the analysis by appending a FOR-clause to the STATS command. Thus the command

```
EAD STATS OF ROWS OF WTLOSS FOR COLS <>1
```

computes descriptive statistics on the rows but excludes column 1 from the calculations (COLS <> 1); the results are shown in Figure 1-4.



STATS OF WTLOSS: WEIGHT OF MALE PATIENTS 7R X 6C

0 STATS	1 -N-	2 -SUM-	3 -MEAN-	4 -VARIANCE-	5 -STDEV-	6 -CI95-
1. ROW 1	4.	1073.	268.25	1128.25	33.58943	53.4072
2. ROW 2	4.	1069.	267.25	1189.583	34.49034	54.83964
3. ROW 3	4.	1039.	259.75	1019.583	31.93092	50.77015
4. ROW 4	4.	998.	249.5	773.6667	27.81486	44.22563
5. ROW 5	4.	959.	239.75	690.25	26.27261	41.77345
6. ROW 6	4.	932.	233.	702.	26.49528	42.1275
7. ROW 7	4.	906.	226.5	811.6667	28.48976	45.29873

Figure 1-4

Notice that each column contains a single measure for each row of WTLOSS (see User's Manual, Section 4, for a complete description of the FOR-clause). However, it is not necessary to exclude explicitly columns that are of nonnumeric data types (e.g., TEXT). This is done automatically.

Setting STATS

If you expect to use these statistics later, you can make a temporary table by naming the resulting table. To do this, use the SET command with the STATS expression. For example

```
SET WTSTATS TO STATS OF ROWS OF WTLOSS FOR COLS <>1
```

This command would make the data in Figure 1-4 into a temporary table named WTSTATS, which can be used later in the session. When you log out the table will be lost. (Note that you can make a permanent table from this temporary table using the MAKE... FROM ... command.)

STATS on a Whole Table

If neither rows nor columns are specified in the DISPLAY STATS command then the statistics will be computed for both the rows and the columns. In this case the displayed table will also include the original data. For example, if you were interested in the differences between patients and how the weight changes over time, the following command

```
SCAN STATS OF WTLOSS FOR COLS <>1
```

will produce Figure 1-5.

The resultant table displays statistics on both rows and columns, and the bottom right-hand diagonal in Figure 1-5 contains the measures computed by taking the whole table as a single sample. For example, the total number of observations in the table, 28, is in cell row "-N-", column "-N-".

<u>Warning</u>: Computing STATS on both rows and columns of a table results in a table containing the original data. Because this table is a temporary table, which is stored in memory rather than on disk, there is some possibility you will run short of space when the data table is large. When you think this may be a problem you should use the ADD STATS command, explained below.

STATS on Arrays

STATS can also be computed on the rows, columns, or both rows and columns of a FLOAT array array in the same manner as with a table. In computing STATS of a FLOAT array, row and column specifications are not necessary, of course. The only restrictions are that you cannot select the data using the FOR-clause and that you cannot add STATS to the array.

STATS OF UTLOSS: WEIGHT OF MALE PATIENTS 13R X 10C

0 STATS	1 PATIENT 1 (LBS)	2 PATIENT 2 (LBS)	3 PATENT 3 (LBS)	4 PATIENT 4 (LBS)	5 -N-	6 -SUM-
1. ROW 1	253.	232.	310.	278.	4.	1073.
2. ROW 2	251.	231.	311.	276.	4.	1069.
3. ROW 3	243.	227.	300.	269.	4.	1039.
4. ROW 4	232.	221.	281.	264.	4.	998.
5. ROW 5	222.	213.	267.	257.	4.	959.
6. ROW 6	215.	206.	260.	251.	4.	932.
7. ROW 7	210.	196.	258.	242.	4.	906.
8. -N-	7.	7.	7.	7.	28.	
9. -SUM-	1626.	1526.	1987.	1837.		6976.
10. -MEAN-	232.2857	218.	283.8571	262.4286		
11. -VARIANCE-	299.2383	184.6667	535.1432	174.9525		
12. -STDEV-	17.29851	13.58921	23.13316	13.22696		
13. -CI95-	16.01864	12.58379	21.42161	12.24834		

STATS OF UTLOSS: WEIGHT OF MALE PATIENTS 13R X 10C

0 STATS	7 -MEAN-	8 -VARIANCE-	9 -STDEV-	10 -CI95-
1. ROW 1	268.25	1128.25	33.58943	53.4072
2. ROW 2	267.25	1189.583	34.49034	54.83964
3. ROW 3	259.75	1019.583	31.93092	50.77015
4. ROW 4	249.5	773.6667	27.81486	44.22563
5. ROW 5	239.75	690.25	26.27261	41.77345
6. ROW 6	233.	702.	26.49528	42.1275
7. ROW 7	226.5	811.6667	28.48976	45.29873
8. -N-				
9. -SUM-				
10. -MEAN-	249.1429			
11. -VARIANCE-		948.6453		
12. -STDEV-			30.80009	
13. -CI95-				11.93237

Figure 1-5

1.4 PROPHET Usage: ADD STATS and DELETE STATS

ADD STATS

You can also append the computed statistics to the original data table by adding the measures to the data table. To do this, use ADD STATS, which will add the statistics to the columns, to the rows, or to both. For example,

```
ADD STATS OF COLS TO PH
```

will add 6 rows of measures to the table PH as shown in Figure 1-6.

PH: PH LEVELS OF 4 PONDS 14R X 4C

0	1 POND 1	2 POND 2	3 POND 3	4 POND 4
1.	7.68	7.69	7.74	7.71
2.	7.7	7.7	7.75	7.71
3.	7.72	7.71	7.77	7.74
4.	7.73	7.73	7.78	7.79
5.	7.73	7.74	7.8	7.84
6.	7.76	7.74	7.81	7.85
7.	7.78	7.78	7.81	7.87
8.	7.8	7.81		7.91
9. -N-	8.	8.	7.	8.
10. -SUM-	61.9	61.9	54.46	62.42
11. -MEAN-	7.7375	7.7375	7.78	7.8025
12. -VARIANCE-	.001619611	.001649039	.0008010864	.005851201
13. -STDEV-	.0402444	.04060836	.02830347	.07649314
14. -CI95-	.03357936	.03388306	.02620938	.06382481

Figure 1-6

<u>Advisories about ADD STATS</u>: Adding the statistics to the table is convenient; however, the computed statistics will <u>not</u> change automatically as the data in the table are changed. You must, therefore, use ADD STATS each time the data are changed in order to update the statistics.

When ADD STATS is performed for a second time you will be warned that some statistics have already been added and asked whether you would like to continue. If you respond "NO" the program will stop; if you respond "YES", it will recompute the statistics. Care must be taken that no column or row of your own data have the same headings as those used by STATS, for they will be written over if they do.

It is not possible to add statistics to a derived column without removing the derivation. You will therefore be asked whether you want to have the derivation (not the data) deleted, or not to have the statistics added to that column. Because it is not statistically correct practice to include in a sample data points whose values intrinsically depend on the values of other data points in the sample, you will be advised if any columns are derived when doing STATS by rows. Normally, in computing STATS by rows on a table that contains many derived columns, judicious use of the FOR-clause specification is recommended.

DELETE
STATS

Statistics that have been added to a table can be deleted with the DELETE STATS command.

```
                     COLUMNS
DELETE STATS  [OF              ] OF table
                     ROWS
```

All the statistics (both row and column) will be deleted when no row or column specification is made.

1.5 Additional Measures

There are several other measures that are often important in providing a more complete statistical description of a sample.

A. Standard error of the mean: SEM

It is often useful to derive a measure of the precision of the estimated mean, μ . If repeated random samples of size n are drawn from a population, the means calculated from each sample will not all be equal, but will vary about the true population mean. The variance of those means will be σ^2/n , where σ^2 is the variance of the original population. From this one can see that the sample means have less variation than the original observations, and that as the sample size gets larger the sample mean becomes a better estimator of the population mean. The estimate of this value (since σ^2 is unknown) is s^2/n , and thus the standard deviation of the mean, more commonly known as the standard error of the mean SEM is

$$\text{SEM} = s_{\bar{x}} = s/\sqrt{n}$$

B. Coefficient of variation: CV%

Often the magnitude of the standard deviation varies directly with the mean. The coefficient of variation is a useful measure of the relative variability of a sample, independent of the scale of the data. The coefficient of variation (CV) is estimated as $s/\bar{x}$ and is usually expressed as a percentage (i.e., 100 $s/\bar{x}$%). One use of the coefficient of variation is to evaluate the relative precision of one experiment compared with another.

C. Median

If the observations of a sample are ranked in order of their magnitude (either ascending or descending) the median is defined as the middle observation. If the number of observations is odd, a unique middle value exists. If the number of observations is even, the median is taken as the average of the two middle observations. That is

$$\text{MEDIAN} = \begin{cases} X_{(n+1)/2} & , \text{ n odd} \\ \dfrac{X_{(n/2)} + X_{(n/2)+1}}{2} & , \text{ n even} \end{cases}$$

If the population is symmetric then the mean and median will be equal and one would expect the sample mean and median to be similar. The median is less sensitive to outliers (unusually high or low data points) than is the mean.

D. Geometric mean: GEOMEAN

The geometric mean is defined as

$$\text{GEOMEAN} = \sqrt[n]{X_1 X_2 \ldots X_n} = e^{\Sigma \ln(X_i)/n}$$

The geometric mean is used when effects are proportional rather than additive. For example, if data in a sample are consistently 18% higher than another sample, (rather than 5 units, say) taking the logarithm of the data will stabilize the variance when the standard deviation in the original scale varies directly with the mean.

E. Minimum and maximum

The smallest (minimum) and largest (maximum) sample points are useful for indicating outliers.

F. Range

The range of the sample is the difference between the sample's maximum and minimum values. As the sample size becomes larger, one would expect the range to increase due to chance alone. In itself the range is not a particularly good descriptor of the sample variation since it is a function of only two values and depends on the sample size.

G. Mid-range

The mid-range is the middle of the range, i.e. (maximum-minimum)/2.

H. Skewness

Another measure that describes the distribution of a population or sample is the skewness. Its chief importance lies in the study of the symmetry of distributions (see Section 2). The distribution is positively skewed if the median is larger than the mean and negatively skewed if the median is smaller than the mean.

The skewness measure, γ_1 is defined for a population as

$$\gamma_1 = \frac{\sum_{i=1}^{n}(X_i-\mu)^3}{N\sigma^3}$$

where N is the size of the population. γ_1 has no units of measurement; it is estimated from the sample data as

$$g_1 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^3}{(n-1)(n-2)s^3/n}$$

I. Kurtosis

The shape ("flatness") of a distribution is measured by the kurtosis. It is defined for a population as

$$\gamma_2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^4}{N\sigma^4} - 3.0$$

and like skewness is dimensionless. It is estimated from the sample data as

$$g_2 = \frac{n(n+1)\sum_{i=1}^{n}(X_i-\overline{X})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Generally, the kurtosis is compared to kurtosis of the normal distribution and hence γ_2 has been corrected by -3 to make the normal distribution $\gamma_2 = 0$. Those distributions with $\gamma_2<0$ are flatter than the normal distributions and those with $\gamma_2>0$ are more peaked (e.g., Student's t-distribution).

1.6 PROPHET Usage: MEASURES

All of the measures described in this section, including those in STATS, are computed by the MEASURES expression. MEASURES is used just as STATS is, e.g., with DISPLAY, PRINT, ADD, or SCAN. For example, you might want to obtain more information about the data on pH levels in four ponds (Figure 1-1).

Figure 1-7 shows the results.

MEASURES OF PH: PH LEVELS OF 4 PONDS 17R X 4C

0 MEASURES	1 POND 1	2 POND 2	3 POND 3	4 POND 4
1. -N-	8.	8.	7.	8.
2. -SUM-	61.9	61.9	54.46	62.42
3. -MEAN-	7.7375	7.7375	7.78	7.8025
4. -VARIANCE-	.001619611	.001649039	.0008010864	.005851201
5. -STDEV-	.0402444	.04060836	.02830347	.07649314
6. -CI95-	.03357936	.03388305	.02620938	.06382481
7. -CI99-	.0497999	.05025029	.03968849	.09465543
8. -SEM-	.01422854	.01435722	.01069771	.02704441
9. -CV%-	.5201214	.5248254	.3637978	.9803671
10. -MEDIAN-	7.73	7.735	7.78	7.815
11. -GEOMEAN-	7.737408	7.737407	7.779955	7.802172
12. -MIDRANGE-	7.74	7.75	7.775	7.81
13. -MINIMUM-	7.68	7.69	7.74	7.71
14. -MAXIMUM-	7.8	7.81	7.81	7.91
15. -RANGE-	.12	.12	.06999999	.2
16. -SKEWNESS-	.5350888	.8680494	-2.153531	-.05194973
17. -KURTOSIS-	-132.5414	-251.1529	1617.792	4.879649

Figure 1-7

ADD MEASURES

MEASURES can be addded to a table in the same way that STATS can. Thus

ADD MEASURES OF COLS OF PH

will add the measures to PH. If STATS had been added to PH previously, MEASURES will recompute them and include the additional measures.

To delete the measures you use the DELETE STATS command described above.

1.7 PROPHET Usage: SELECT STATS

You may find that the set of measures computed by STATS does not provide the most adequate description for some kinds of data. It might be that you want to compute the median or the coefficient of variation, for example, but the full set of statistics in MEASURES is more than you need. PROPHET allows you to make a tailored set of measures which you would like the STATS expressions to compute.

The SELECT STATS quiz will dislay all seventeen measures available and indicate with asterisks which measures are presently being used in STATS. You will then be asked to specify (by listing indices) the measures that STATS should compute. This list of indices will be used, in the order specified, to set the measures which will be computed each time STATS is called in the session. The list will be unchanged until you log out of the session, or until you use the SELECT STATS command again.

For example, if you were concerned about outliers, you might want to include MINIMUM and MAXIMUM along with the other statistics. After the command

```
SELECT STATS
```

is typed you will see the full set of measures displayed. Figure 1-8 below shows what would appear on the screen after you have selected the statistics by typing the list of indices: "1, 2, 3, 5, 6, 10, 11". The updated list for STATS is shown on the right hand side in the order specified.

You will also be asked how many decimal places to which you would like to round off the statistical output; this must be a positive fixed number. The columns will be formatted to show that many digits to the right of the decimal point. The data themselves will be unaffected. When the ADD command is used to compute the measures, columns containing the original data will not be formatted.

```
>SELECT STATS<

THE FULL SET OF STATISTICAL MEASURES AVAILABLE IS SHOWN BELOW.
THOSE CURRENTLY SELECTED ARE STARRED.

TO LEAVE THEM UNCHANGED, JUST TYPE <GO>.

TO RETURN TO THE SYSTEM STANDARD, TYPE "STANDARD" OR "S".

SELECT THE MEASURES YOU WANT BY TYPING IN THEIR NUMBERS,
SEPARATED BY A SPACE OR A COMMA.
?1,2,3,5,6,10,11<

DO YOU WANT THE STATISTICS ROUNDED OFF?Y<
  HOW MANY DECIMAL PLACES DO YOU WANT?3< >
```

```
 ________________  _______________  __________
  1 | *  N          13 |   MINIMUM    1 | N
  2 | *  SUM        14 |   MAXIMUM    2 | SUM
  3 | *  MEAN       15 |   RANGE      3 | MEAN
  4 | *  VARIANCE   16 |   SKEWNESS   4 | STDEV
  5 | *  STDEV      17 |   KURTOSIS   5 | CI95
  6 | *  CI95                          6 | MEDIAN
  7 |    CI99                          7 | GEOMEAN
  8 |    SEM
  9 |    CV%
 10 |    MEDIAN
 11 |    GEOMEAN
 12 |    MIDRANGE
```

Figure 1-8

Standard STATS

To return to the standard set of measures for STATS and no formatting of output, use the following command:

SELECT STATS TO STANDARD

Note that the set for MEASURES is unaffected by any selection within the SELECT STATS routine, and MEASURES commands can be issued without having any effect on the set of STATS selected.

1.8 PROPHET Usage: LOGSTATS and LOGMEASURES

It is not unusual in biostatistical applications that the analysis should be performed on the logarithm of the data. It is appropriate, for example, when the coefficient of variation is constant, i.e., the standard deviation in the original scale varies directly as the mean. This occurs when effects are proportional, not additive. To have the natural logarithm of the data taken before the computations are made, use the LOGSTATS and LOGMEASURES expressions instead of STATS and MEASURES.

PRESSURE: CHANGE IN BLOOD PRESSURE (MM HG) 12R X 4C

0 DOG NAME	1 0.005 MICROGM/ KILOGM	2 0.01 MICROGM/ KILOGM	3 0.025 MICROGM/ KILOGM	4 0.05 MICROGM/ KILOGM
1. GANDALF	5.	10.	18.	29.
2. TIGGIE	4.	12.	17.	30.
3. LUCIOUS	1.	6.	14.	26.
4. FRODO	2.	9.	16.	32.
5. ARISTOTLE	1.	8.	19.	25.
6. -N-	5.	5.	5.	5.
7. -SUM-	3.688879	10.85592	14.07967	16.7112
8. -MEAN-	.7377759	2.171183	2.815934	3.34224
9. -STDEV-	.7534901	.2592035	.1178104	.1020461
10. -CI95-	.9367794	.3222558	.1464683	.1268692
11. -MEDIAN-	.6931472	2.197225	2.833213	3.367296
12. -GEOMEAN-	2.091279	8.768655	16.70877	28.28241

Figure 1-9

In Figure 1-9, selected statistics have been added to the columns of the table PRESSURE using

ADD LOGSTATS OF COLS OF PRESSURE

The data consist of the change in blood pressure of five dogs in response to 4 drug dose levels. All the measures (except the geometric mean) are computed on the natural logarithms of the data. When a row or column contains negative or zero values, LOGSTATS and LOGMEASURES will not be computed for that row or column, nor will they be computed for the data taken as a whole (i.e., totals).

1.9 PROPHET Usage: Individual Measures

In addition to computing STATS and MEASURES on all the rows or columns of a table it is possible to compute certain single measures on a single row or column. Thus each of the commands

```
TYPE MEAN OF COL 2 OF PH
DISPLAY MEDIAN OF ROW 6 OF PH
SET X TO STDEV OF COL 3 OF PH
```

will return a single FLOAT number. If the row or column specification is omitted the whole table is taken as a single sample. The measures available are:

SUM, MEAN, VARIANCE, STDEV, MEDIAN, MINIMUM, MAXIMUM

Individual measures can also be computed on the rows or columns of FLOAT and FIXED array arrays and on FLOAT and FIXED arrays.

1.10 PROPHET Usage: Graphs

Once STATS or MEASURES have been computed it is often useful to make a graph to visualize the data. From MEASURES computed in Figure 1-7 we can make an isograph showing the means in curve 1 and confidence intervals about them.

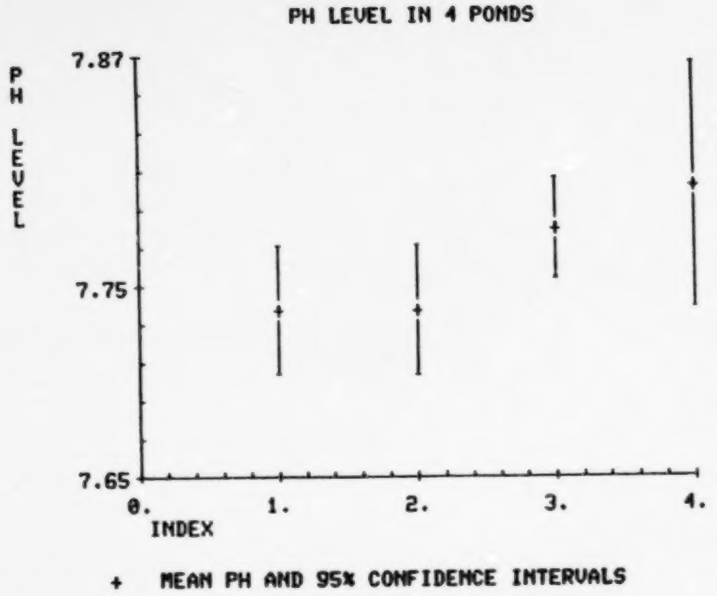


Figure 1-10

It is also useful to add two additional curves containing the minimum and maximum values for each sample in order to check whether differences in the means may be attributed to outlying points and not to differences in the major cluster of data.

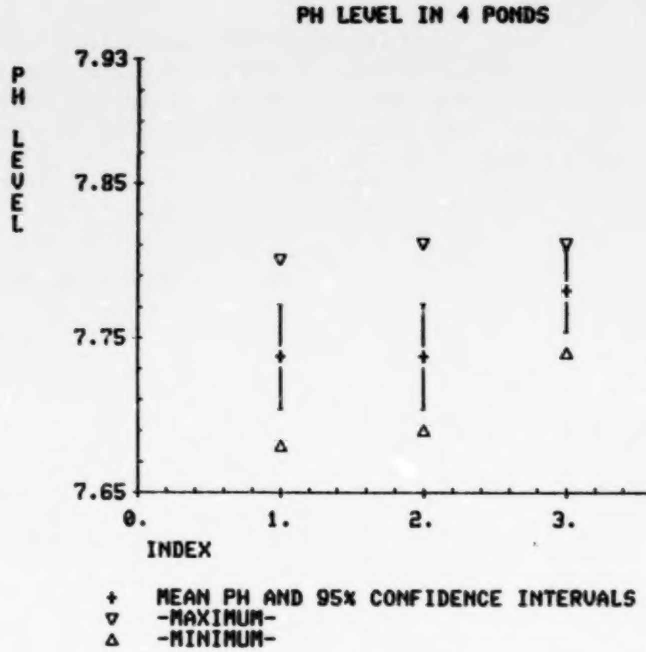


Figure 1-11

These are just a few of the possible ways you might want to display the data.

One wants to make the graph so that it summarizes the data set without losing any of its unusual characteristics.

2. Distributions

2. Distributions

SECTION 2

DISTRIBUTION TESTING

2.1 Introduction

In Section 1 it was pointed out that the observations of a sample range over a series of values, arranged in some fashion about a measure of central tendency, such as the mean. However, the actual distribution of observations about the central measure may vary considerably among samples.

Studying the distribution may reveal several peculiarities in the data, such as unusual trends or obvious outliers, that might otherwise be missed. Furthermore, the reliability, and in many instances the validity, of a wide range of statistical tests relies heavily on the nature of the distribution. A certain category of statistical procedures, known as _parametric_ tests, almost always assumes that the data are distributed according to the _normal_ distribution (see below). However, even the nonparametric tests often make assumptions about the symmetry of the data and/or the equality of dispersions.

One basic way to view how the observations are distributed is to divide the data points into intervals and count the number of observations found in each interval. The resulting table is a frequency distribution that can be displayed graphically as a histogram (a bar graph on ratio or interval data).

2.2 PROPHET Usage: Making Histograms

In Figure 2-1 the table HEIGHT shows the height (in inches) of the first seventy graduate students in a biostatistics course (Zar,p. 79). To construct a frequency table HEIGHTDIST from HEIGHT, we use the following command:

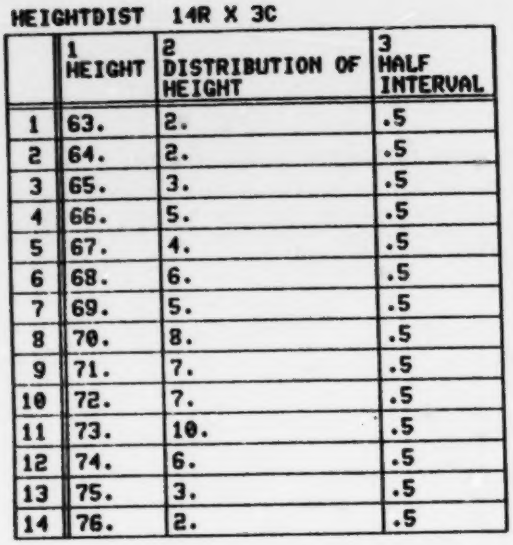
```
MAKE TABLE HEIGHTDIST AS DISTRIBUTION OF COLUMN 1 OF HEIGHT
```

HEIGHT: HEIGHTS OF 70 GRADUATE STUDENTS 70R X 2C

	1 INCHES		1 INCHES		1 INCHES		1 INCHES
1	72.	20	65.	39	71.	58	74.
2	73.	21	72.	40	73.	59	68.
3	73.	22	68.	41	74.	60	73.
4	73.	23	73.	42	73.	61	66.
5	68.	24	75.	43	70.	62	66.
6	67.	25	64.	44	66.	63	70.
7	71.	26	69.	45	74.	64	75.
8	76.	27	72.	46	71.	65	74.
9	68.	28	69.	47	70.	66	70.
10	64.	29	66.	48	67.	67	68.
11	76.	30	70.	49	74.	68	63.
12	71.	31	69.	50	74.	69	63.
13	69.	32	70.	51	65.	70	69.
14	72.	33	70.	52	71.		
15	73.	34	73.	53	65.		
16	75.	35	71.	54	72.		
17	70.	36	72.	55	66.		
18	67.	37	72.	56	67.		
19	68.	38	73.	57	71.		

Figure 2-1

The first column of HEIGHTDIST (Figure 2-2) contains the values at the center of the intervals; the third column contains the width of the half intervals to the left and right. Since no specification was made for the intervals, intervals were created so that each one contains just one different data value. All occurrences of that value were then counted and recorded in the second column. For example, row 7 indicates that there are five heights in the range of 68.5 to 69.5 inches.



HEIGHTDIST 14R X 3C

	1 HEIGHT	2 DISTRIBUTION OF HEIGHT	3 HALF INTERVAL
1	63.	2.	.5
2	64.	2.	.5
3	65.	3.	.5
4	66.	5.	.5
5	67.	4.	.5
6	68.	6.	.5
7	69.	5.	.5
8	70.	8.	.5
9	71.	7.	.5
10	72.	7.	.5
11	73.	10.	.5
12	74.	6.	.5
13	75.	3.	.5
14	76.	2.	.5

Figure 2-2

You can specify the interval sizes in either of two forms. One way is to specify a fixed interval width and either the starting value for the first interval or the ending value for the last interval. The second way is to specify explicitly the endpoints of the intervals in a column of a table or in a float array. The intervals need not have the same widths. For example, both the commands below are equivalent to the command above and will also produce the table HEIGHTDIST in Figure 2-2:

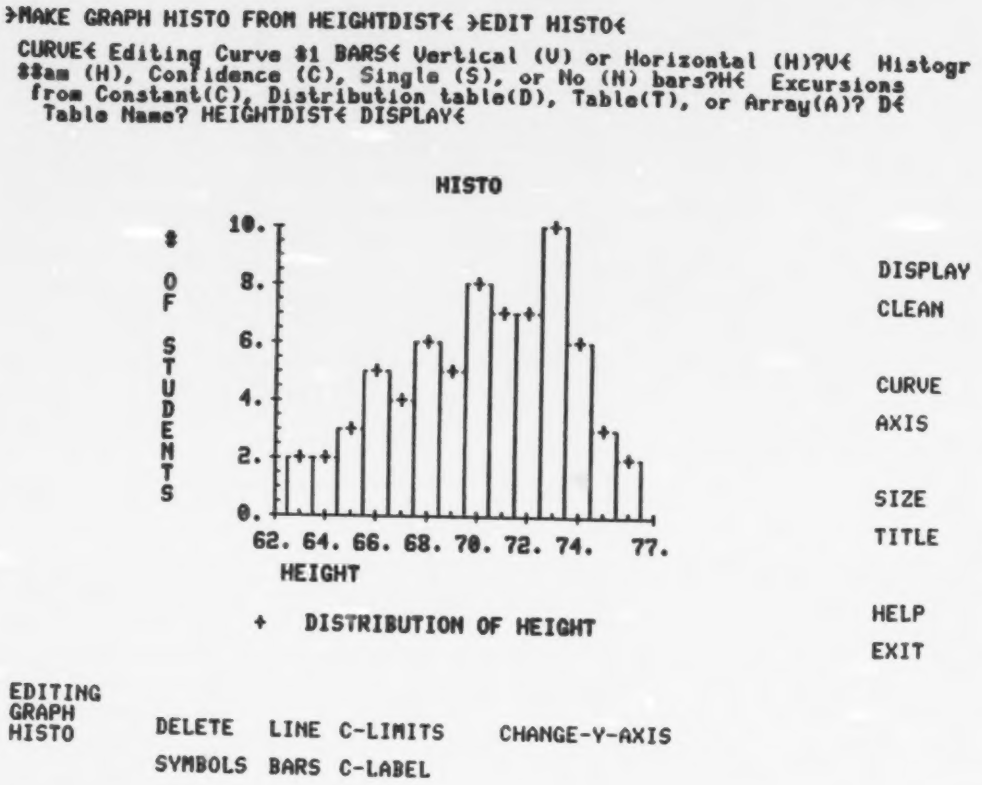
```
MAKE TABLE HEIGHTDIST AS DISTRIBUTION OF COLUMN 1 OF HEIGHT
    FOR INTERVALS OF 1. STARTING AT 62.5

MAKE TABLE HEIGHTDIST AS DISTRIBUTION OF COLUMN 1 OF HEIGHT
    FOR INTERVALS OF COLUMN 2 OF HEIGHT
```

where the second column of HEIGHT contains 62.5, 63.5, 64.5, ..., 76.5. Usually, the intervals should be of a constant size for making a histogram in order to facilitate comparison among the intervals. But it is possible to vary the widths of the intervals in a distribution table. For example,

```
MAKE TABLE HEIGHTDIST AS DISTRIBUTION OF COLUMN 1 OF HEIGHT
  FOR INTERVALS [62.5, 65.5, 67.5, 69.5, 71.5, 73.5, 76.5]
```

will make a distribution table where the first and next to last intervals are larger than the other intervals.

A histogram of the data can be made from a distribution table HEIGHTDIST as shown in Figure 2-3.



```
>MAKE GRAPH HISTO FROM HEIGHTDIST< >EDIT HISTO<
 CURVE< Editing Curve #1 BARS< Vertical (V) or Horizontal (H)?V<  Histogr
 ##am (H), Confidence (C), Single (S), or No (N) bars?H<  Excursions
  from Constant(C), Distribution table(D), Table(T), or Array(A)? D<
   Table Name? HEIGHTDIST< DISPLAY<
```

HISTO
# OF STUDENTS
10.
8.
6.
4.
2.
0.
62.
64.
66.
68.
70.
72.
74.
77.
HEIGHT
+ DISTRIBUTION OF HEIGHT
DISPLAY
CLEAN
CURVE
AXIS
SIZE
TITLE
HELP
EXIT
EDITING GRAPH HISTO
DELETE
LINE
C-LIMITS
CHANGE-Y-AXIS
SYMBOLS
BARS
C-LABEL

Figure 2-3

The first command in Figure 2-3, "MAKE GRAPH HISTO", makes a graph of the column 1 vs. column 2 of HEIGHTDIST. To make the curve into a histogram you must then edit the graph. First specify CURVE and then BARS by pointing to the light buttons. Then you will be prompted for the remaining items. You would specify Vertical bars, Histogram, and Excursions from a Distribution table named HEIGHTDIST. For more details on bar graphs see Section 5 of the User's Manual.

2.3 Normal Distribution, Skewness and Kurtosis

Of a large number of theoretical distributions of various physical phenomena, the most important is the normal distribution; its properties form the basis of most parametric statistical analyses.

Suppose thousands of samples of size ten are drawn from a population with an arbitrary distribution. If you were to make a histogram of the means of these samples you would get a distribution that will have a shape similar to the distribution of the original population. Now suppose you draw thousands of samples of size ten thousand from the same population. This time a histogram of the means will look more like a normal distribution than when the sample size was ten. The histogram of the means will become closer to a normal distribution as the size of the sample becomes larger.

This is stated more precisely in the Central Limit Theorem: If the distribution of a population (not necessarily normal) has a finite variability, then the sampling distribution of the means of random samples of size n drawn from the population will approach normality as the sample size approaches infinity. Thus, if the sample size is large enough one can compare sample means using normal probability theory. How "large" is large enough depends on the distribution of the original population. The more sharply it differs from a normal distribution, the larger the sample size must be for the means to follow a normal distribution.

The histogram in Figure 2-3 is the most useful way of finding anomalies in the distribution of the data. For parametric tests one wants to see the tops of the bars follow a symmetric "bell-shape" curve, which indicates a distribution similar to a normal distribution. When the curve does not follow the bell model, it may be that the data should be transformed. For example, if the data are highly skewed (that is, higher bars concentrated on one side of center) then taking a transformation such as the logarithm, square root, or square might help to bring the distribution closer to normality. The existence of two distinct clusters of higher bars might suggest that the data actually belong to two distinct populations.

While the histogram provides a good initial idea of whether the data follow a normal distribution, a statistical test is ordinarily useful. PROPHET has two tests for assessing whether there is evidence that the data are nonnormal. When the sample size is less than or equal to 50, the Wilk-Shapiro test is used; otherwise the D'Agostino test is used.

A. Wilk-Shapiro (W-S) test for normality

The W-S test provides an index or statistic to test the hypothesis that a sample comes from a population that is normally distributed. This statistic is an effective measure of normality even for small samples ($n \leq 50$) against a wide spectrum of nonnormal alternatives.

To compute the value of W, given a sample of size n, the procedure is as follows:

- Order the observations by value:

$$X_1 \leq X_2 \leq X_3 \leq \ldots \quad X_n$$

- Compute the sum of squares:

$$SS = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Compute the weighted sum of differences:

$$b = \sum_{i=1}^{k} a_i (X_{n-i+1} - X_i)$$

  where the values of the coefficients, a, are taken from a table of constants and either $k=n/2$ if n is even, or $k=(n-1)/2$ if n is odd.

- Compute the W statistic: $W = b^2/s^2$

- W is compared to the critical values at the .05 and .01 levels. Small values of W are significant, i.e., they indicate nonnormality.

B. D'Agostino's test for normality

For larger sample sizes ($n>50$) PROPHET performs D'Agostino's test for normality. As in the Wilk-Shapiro test, the observations are ordered in ascending order and the sum of squares, SS, is computed. D'Agostino's D-statistic is computed as:

$$D = \frac{\sum_{i=1}^{n} (i - (n + 1)/2) X_i}{\sqrt{n^3 SS}}$$

The computed value of D is compared to a table of critical values.

C. Remarks on the tests for normality

The null hypothesis for the Wilk-Shapiro and D'Agostino tests is that the data come from a normally distributed population. If the null hypothesis is not rejected, however, remember that the test does not prove that the population is normally distributed; it simply indicates that there is not sufficient evidence to reject that assumption.

These tests for normality are general tests; they determine whether a sample distribution deviates from normality but not how it deviates. If the population is not normally distributed, it is often desirable to know what characteristic of the distribution is different from normal distributions. In fact, when you know this, then you can often transform the data into a normal distribution.

D. Skewness and kurtosis

In the same way that the means of samples from a single population vary from sample to sample, the calculated skewness and kurtosis vary. The distributions of the sample skewness, g_1, and kurtosis, g_2, depend upon the distribution of the original population. Populations that are highly skewed will lead to different distributions for skewness and kurtosis than will populations that are symmetric. If the population is normal then the distributions of g_1 and g_2 are known to be normal when the sample size, n, is large. The mean and variance for skewness are:

$$\text{Mean}\,(g_1) = 0.0$$

$$\text{Variance}\,(g_1) = \frac{6n(n-1)}{(n-2)(n+1)(n+2)}$$

The mean and variance for kurtosis are:

$$\text{Mean}\,(g_2) = 0.0$$

$$\text{Variance}\,(g_2) = \frac{24n(n^2-1)(n-1)}{(n-3)(n-2)(n+1)(n+3)(n+5)}$$

To test whether the measures of skewness and kurtosis differ significantly from the values that would be obtained if the data were normally distributed, you begin by assuming the two null hypotheses that both g_1 and $g_2 = 0$. The calculated measures divided by the square roots of their variances are compared to a normal distribution with mean 0 and variance 1. If either statistic differs significantly from 0, the corresponding null hypothesis is rejected.

For smaller sample sizes, the distribution of skewness is not normal, but there are tabulated critical values that may be used to test the null hypothesis. For the test on skewness, these tabulated values are used for $8<n<500$. When the calculated value exceeds the critical value, there is significant evidence to reject the null hypothesis. Geary developed a test criterion based on kurtosis when $11<n<1001$. The test statistic is not the standard formula for kurtosis but the following:

$$a = \frac{\sum_{i=1}^{n}|X_i - \overline{X}|}{n} \Big/ \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/n}$$

This is the average absolute difference from the mean divided by the "standard deviation". When the data are normally distributed the expected value of a is 0.7979. By comparing the computed value of a with the tabulated value, you can test whether a is greater than expected by chance alone.

Since most parametric tests are quite robust, slight to moderate deviations from normality will probably not invalidate the test. But if either skewness or kurtosis is significantly different from zero then perhaps a transformation on the data should be made. (Zar, pp. 182-89). If a transformation does not correct the problem then you should use either a parametric test designed for data from a population with a distribution your data is from or a nonparametric test. However, even nonparametric tests, which make no assumption about the type of distribution the data are from, will sometimes assume symmetry.

2.4 PROPHET Usage: TEST NORMALITY and TEST NORMALSHAPE

Testing for Normality

To test whether a sample comes from a population of normal data you type the command

TEST NORMALITY

You are asked for the name of the table, whether it is organized by rows or columns, and the row or column number of the sample. If the sample size is less than or equal to 50, the Wilk-Shapiro test will be performed; otherwise, D'Agostino's test will be performed. Figure 2-4 shows the results of the test for the data in Figure 2-1. Since n>50 D'Agostino's test is performed.

```
>TEST  NORMALITY

NAME OF TABLE  ?HEIGHT<
DATA FROM ROWS OR COLUMNS (R OR C)?C<
COLUMN NUMBER FOR GROUP 1?1<

MEAN = 70.17143   VARIANCE = 10.95567   N = 70

NULL HYPOTHESIS  H0: THE POPULATION FOLLOWS THE NORMAL DISTRIBUTION.

D'AGOSTINO'S TEST: CALCULATED D = .2843672
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.
```

Figure 2-4

Shape of Distribution

To test the data for skewness and kurtosis you use the command

TEST NORMALSHAPE

You are asked for the name of the data table and the row or column number of the sample. The results of the test for the data in Figure 2-1 are shown in Figure 2-5.

```
>TEST NORMALSHAPE<


NAME OF TABLE  ?HEIGHT<
DATA FROM ROWS OR COLUMNS (R OR C)?C<
COLUMN NUMBER FOR SAMPLE 1?1<

DO YOU WISH TO TEST ON THE BASIS OF SKEWNESS (S), KURTOSIS (K) OR BOT-
H (B)?B<

TEST STATISTIC: SKEWNESS
EXPECTED VALUE = 0.0
SAMPLE VALUE = -.3453076
INFERENCE: DEPARTURE FROM NORMALITY IS NOT SIGNIFICANT AT P>0.05


TEST STATISTIC: GEARY'S MEASURE OF KURTOSIS
EXPECTED VALUE = 0.7979
SAMPLE VALUE = .2521672
INFERENCE: DEPARTURE FROM NORMALITY IS SIGNIFICANT AT P<0.05
```

Figure 2-5

2.5 Testing Other Distributions

There are several other distributions which have important theoretical properties and are useful for describing various kinds of data encountered in biostatistical research. The Kolmogorov-Smirnov distance can be used to test a sample against a normal, binomial, uniform, chi-square, Poisson, Student's t, or F distribution.

The rationale behind this test is that the cumulative distribution of the sample should not diverge significantly from the cumulative distribution of the hypothesized distribution. It is defined at X_i as the probability of observing a value less than X_i. It is a nondecreasing function of X and usually follows an S shape curve starting at 0 for the smallest value of X and ending at 1 for the largest value of X.

The graph in Figure 2-6 was produced from the distribution data of Figure 2-2 by continuously accumulating the number of observations, and finally dividing by 70, the total number of observations. The curve is the theoretical cumulative normal distribution curve with mean = 70 and variance = 11. The Kolmogorov-Smirnov distance is the maximum vertical difference between the two curves.

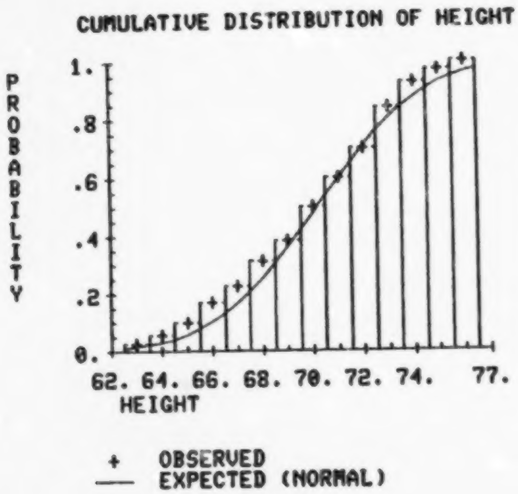


Figure 2-6

The Kolmogorov-Smirnov D statistic is computed by comparing the maximum vertical distance between the observed cumulative distribution of the sample with the expected cumulative distribution at each value of X. If the distributions are not the same, the accumulation will occur at different rates and at some point there will be a large discrepancy between them. The D statistic is computed as

$$D = \begin{cases} \sup_i |F(X_i) - (i - 0.5)/n| \,, & \text{if X is continuous} \\ \max_i |F(i) - \sum_{j=1}^{i} f_j| \,, & \text{if X is discrete} \end{cases}$$

where F(X) is the theoretical cumulative density function.

2.6 PROPHET Usage: TEST DISTRIBUTION

To test the distribution of a sample, type the command

```
TEST DISTRIBUTION
```

You are asked for the name of the table, whether it is organized by rows or columns, and the row or column number of the sample. You are then asked for the name of the distribution to which you wish to compare the sample; valid replies are NORMAL, BINOMIAL, UNIFORM, CHISQUARE, POISSON, T, and F. (You need type only the first letter).

Depending upon the distribution you specify, PROPHET will then ask for other information about the distribution, as shown in Table 2-1. If you do not wish to specify these parameters, simply type <GO>, in which case these parameters will be estimated by the test procedure. The estimated parameters are then displayed, followed by the Kolmogorov-Smirnov test statistic and the inference statement regarding the null hypothesis.

Table 2-1

Parameter Values asked for in TEST DISTRIBUTION

Distribution	Parameter 1	Parameter 2
Normal	Mean	Variance
Binomial	Number of Trials	Probability of Success
Uniform	Lower Bound	Upper Bound
Chi-squared	Degrees of Freedom	(not applicable)
Poisson	Mean	(not applicable)
t	Degrees of Freedom	(not applicable)
F	Numerator Degrees of Freedom	Denominator Degrees of Freedom

PHBLOOD: MEANS OF 3 BLOOD SAMPLES 20R X 1C

	1 PH		1 PH
1	6.702	11	6.725
2	6.723	12	6.7005
3	6.713	13	6.69
4	6.688	14	6.7225
5	6.6905	15	6.692
6	6.725	16	6.7035
7	6.717	17	6.7085
8	6.721	18	6.7025
9	6.735	19	6.7165
10	6.707	20	6.682

Figure 2-7

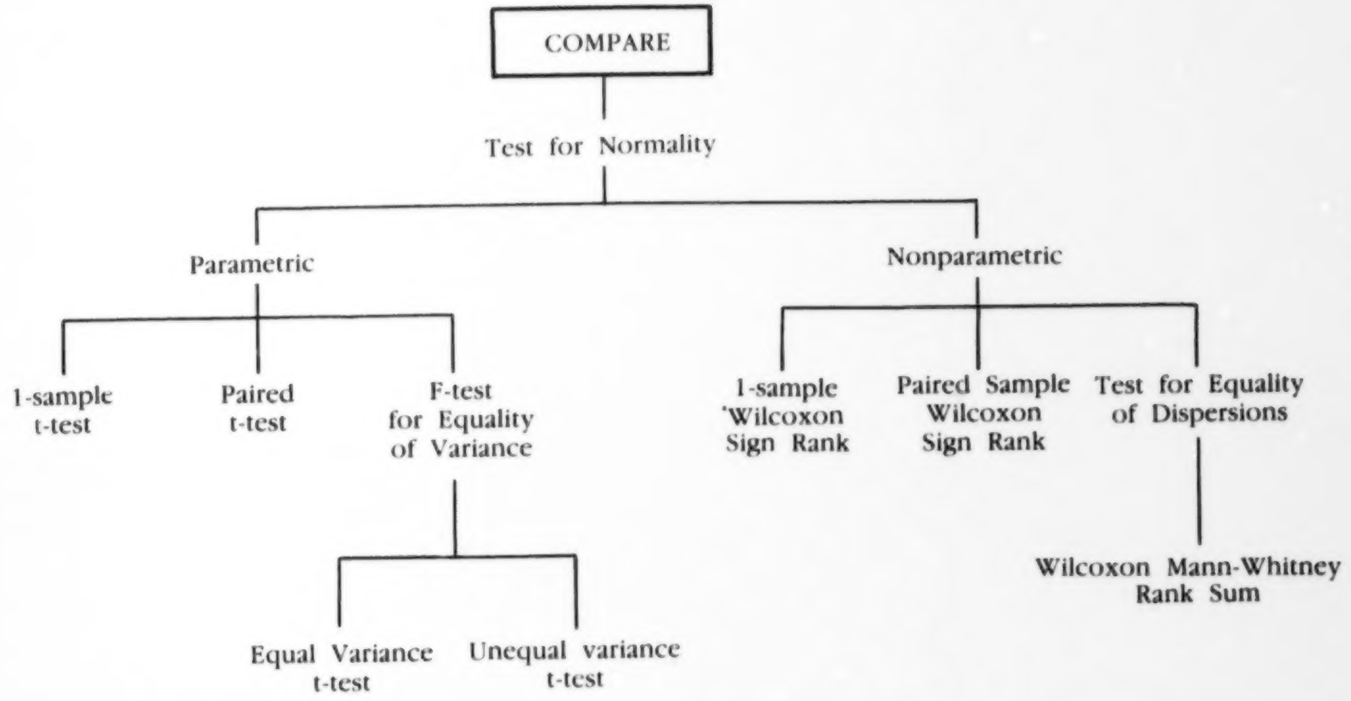
For example, the pH level was measured in three blood samples which were taken from each of twenty experimental subjects. The average for each subject is shown in Figure 2-7.

The TEST DISTRIBUTION command was used to test whether the means come from a population following the Student's t-distribution. The results are shown in Figure 2-8. Since the means were calculated from samples of size three, 2 degrees of freedom was specified.

```
>TEST DISTRIBUTION<

NAME OF TABLE  ?PHBLOOD<
NAME OF DISTRIBUTION TO BE COMPARED TO?T<
TYPE <GO> IF YOU DO NOT WISH TO SPECIFY A PARAMETER.

DEGREES OF FREEDOM =?2<
NULL HYPOTHESIS  H0: THE POPULATION FOLLOWS THE T DISTRIBUTION
MEAN =6.708225,  VARIANCE =.0002208509,  N =20

DEGREES OF FREEDOM =2.

KOLMOGOROV'S D STATISTIC =.96415

INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01. >
```

Figure 2-8

3. COMPARE: I


COMPARE
Test for Normality
Parametric
Nonparametric
1-sample t-test
Paired t-test
F-test for Equality of Variance
Equal Variance t-test
Unequal variance t-test
1-sample 'Wilcoxon Sign Rank
Paired Sample Wilcoxon Sign Rank
Test for Equality of Dispersions
Wilcoxon Mann-Whitney Rank Sum

Section 3

ONE- AND TWO-SAMPLE HYPOTHESIS TESTING

3.1 Statistical Hypothesis Testing

In the Introduction we briefly discussed the notion of statistical hypothesis testing. In general, one performs tests of significance on samples so that inferences about population parameters can be drawn. These tests allow an investigator to attach a measure of confidence to the observations and resultant conclusions. Usually one compares the strength or validity of observed or possible differences to the alternative possibility that these differences are likely to have occurred by random chance.

There is a wide variety of statistical procedures for hypothesis testing, and choosing the appropriate procedure is of first importance. The test must first be suitable to the distribution of the data (normal or nonnormal), to the type of data, and to the number of comparisons. Often more than one test may be valid. In these cases, you should consider the consequences of committing a Type II error as opposed to a Type I error (see Introduction). Finally, even a correct statistical test procedure is not infallible, and judicious interpretation of any result, as opposed to blind acceptance of the stated inference, is fundamental to any statistical analysis.

3.2 Parametric and Nonparametric Methods

The appropriateness of any statistical test rests on certain underlying theoretical assumptions about the sample and population. For example, many test procedures rely on the Student's t-distribution, and it is assumed that the samples were obtained from populations whose distributions are normal. These tests fall into the general category of methods known as parametric procedures.

By contrast, other statistical methods do not use any estimation of the mean, variance, or any other population parameter. Hypotheses concerning these parameters are not considered in these testing procedures, which are known as nonparametric procedures. As they make no assumptions about the distribution of the sampled populations, they are also known as distribution-free tests. Ordinarily, you should use nonparametric procedures when the assumption of normality is clearly violated.

Ordinal scale data are usually analyzed with nonparametric procedures. Although there is nothing in the theoretical basis of parametric tests that requires interval or ratio scale data, it is often the case that ordinal scale data deviate significantly from normality.

There are, however, situations when either approach may be applicable. In these cases the parametric test is usually more powerful, i.e., there is a greater probability of committing a Type II error with the analogous nonparametric procedure.

When the data are discovered to be significantly nonnormal, it is often possible to employ a mathematical transformation (e.g., logarithmic, square root, arcsine) which will cause the resulting distribution to be normal. You can do this very easily in PROPHET by using derived columns. Transformation is a very useful procedure; in fact, some statisticians argue that using parametric procedures on transformed data is preferable to the use of nonparametric tests on the original data. The parametric tests, in fact, tend to be more efficient in detecting departures from the null hypothesis.

In PROPHET data can be tested for normality by the Wilk-Shapiro, D'Agostino, or Kolmogorov-Smirnov tests (see Section 2).

3.3 PROPHET Usage: COMPARE

A variety of procedures in PROPHET perform one- and two-sample tests of significance. All are accessed through the COMPARE command. This command sets up a dialogue that enables you to select which procedure to use. You are prompted for information such as the data table name, how many samples are to be compared, whether the data are organized by rows or columns, and whether you wish to use parametric or nonparametric tests. In certain instances a test will automatically be performed to guide you in selecting the appropriate test to use next (for example, the test for normality and the test for equality of variance). Figure 3-1 shows a typical COMPARE dialogue, and the meaning of each question is described below. Several uses of the COMPARE command will be illustrated in Sections 3.4 and 3.5.

```
ARE YOU COMPARING 1, 2, OR MORE THAN 2 GROUPS (1, 2, OR M)?2←

DO THE DATA COME FROM 1 OR 2 TABLES (1 OR 2)?2←

NAME OF TABLE 1?IQ←
DATA FROM ROWS OR COLUMNS (R OR C)?C←
COLUMN NUMBER FOR GROUP 1?1←

NAME OF TABLE 2?IQSCORE←
COLUMN NUMBER FOR GROUP 2?1←

ARE YOU PERFORMING PARAMETRIC TEST (P), NONPARAMETRIC TEST (N),
OR ARE YOU UNDECIDED (U)?U←

ARE THE DATA PAIRED ?N←

POPULATION COLUMN 'A-12' ARE NORMAL AT P>0.05 BY W-STATISTIC.
POPULATION COLUMN 'IQ(1)' ARE NORMAL AT P>0.05 BY W-STATISTIC.


NULL HYPOTHESIS  H0: THE POPULATION VARIANCES ARE EQUAL.
F = 1.103   DF(N/D) = 9/18   P = .408
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.
<HIT GO TO CONTINUE>
```

Figure 3-1

ARE YOU COMPARING 1, 2 OR MORE THAN 2 GROUPS (1, 2 OR M)?

If you are comparing <u>one</u> sample to a hypothetical value, or standard, you enter 1. If the data are normal, the one-sample t-test will be performed. If the data are nonnormal the nonparametric Wilcoxon one-sample test will be used. In either case you are then quizzed for the standard value.

If you are comparing two samples, you can use data for both samples from the same table, or from two separate tables. Thus you enter "1" or "2" in response to the question "DO YOUR DATA COME FROM 1 OR 2 TABLES?". As the roadmap facing page 3-1 shows, three forms of the two-sample t-test are available: paired, unpaired with equal variances, and unpaired with unequal variances. There are two nonparametric two-sample tests; the Wilcoxon sign rank test for paired data, and the Wilcoxon-Mann-Whitney test for unpaired data. Note that you never have to specify the test by name. The correct choice is made on the basis of your response to COMPARE's questions.

NAME OF TABLE?

You enter the name of the table containing the data. If your data come from two tables, you are asked to supply two names.

DATA FROM ROWS OR COLUMNS (R OR C)?

If the individual samples are contained in columns, enter "C"; if the samples are in rows, you enter "R". COMPARE then asks for the column (or row) number for each sample. Note that the columns must be of type "NUMBER".

ARE YOU PERFORMING PARAMETRIC TEST (P), NONPARAMETRIC TEST (N), OR ARE YOU UNDECIDED (U)?

If you know in advance that you wish to perform either parametric or nonparametric tests, you enter either "P" or "N". Otherwise you enter "U", in which case the test for normality will be made and the results presented. If the data are found to be significantly nonnormal you are given the option of proceeding with either parametric or nonparametric procedures, or aborting.

ARE THE DATA PAIRED?

If you are comparing sets of paired observations you respond "YES", otherwise you respond "NO".

When the data are not paired, either the F-test for equality of variances (parametric) or the Ansari-Bradley test for equality of dispersions (nonparametric) will be performed. If the F-test indicates that the variances are not equal you are asked if you wish to use the unequal variance t-test. If you respond "YES" this test will be performed. Otherwise the equal variance t-test will be used.

Repeating the Test

There is a further option available in COMPARE which facilitates the analysis of many independent sets of data. Frequently, an investigator has a number of data sets to be analyzed identically. For example, an experiment might be designed to test the responses in blood pressure, cardiac output, and heart rate of a group of paired subjects given a certain drug.

In this situation, the same test procedure, e.g., the paired t-test, would be repeated three times to analyze the three sets of observations. To do this in COMPARE, you respond to the first question asked, "ARE YOU REPEATING THE SAME TEST PROCEDURE ON SEVERAL SETS OF DATA?", with "YES", and then enter '3'. As a result, COMPARE then asks general questions such as "ARE THE DATA PAIRED?", "ARE THE DATA ORGANIZED BY ROWS OR COLUMNS?" only once, because these answers are assumed to be the same for each of the three sets of data.

This use of repeated tests should not be confused with multisample comparisons, where more than two data sets of the same response measure are to be compared. (See Section 4.)

3.4 Parametric Tests for One and Two Samples

A. One-sample Test

Often it is of interest to compare the mean of a sample to a known standard value, m. The one-sample t-test is used for this purpose. You begin by stating the null hypothesis that the population from which the sample was drawn has a mean = m.

If the population variance, σ^2, were known, then the statistic

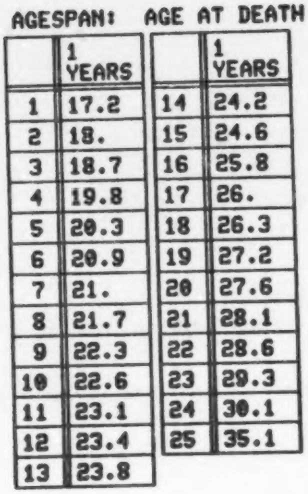
$$t = \frac{\overline{X} - m}{\sigma/\sqrt{n}}$$

is known to follow the normal distribution. However, since σ^2 is rarely known, the sample variance s^2 is substituted in the above expression and the resulting statistic is

$$t = \frac{\overline{X} - m}{s/\sqrt{n}}$$

This statistic is known to follow Student's t-distribution, whose shape varies as a function of the sample size n. In the limiting case, as n approaches infinity, s^2 will approach σ^2 and the t-distribution approaches the normal distribution. With smaller values of n, the shape of the t-distribution becomes more peaked, with the area under the tails (extremities) of the distribution becoming larger. Since the t-distribution is a function of n, an additional parameter, known as the degrees of freedom (DF), is needed for any test using the t-distribution.

The data in Figure 3-2 represent the age (in years) at death of 25 horses (Zar, p. 89). We might test the null hypothesis (H0) that the mean age at death is not different from 22 years. That is, we want to know if we can infer from this sample that the horses were selected from a sample population of horses whose life span equals 22 years. The mean of the data is 24.23 years, and the SEM=0.85 years.

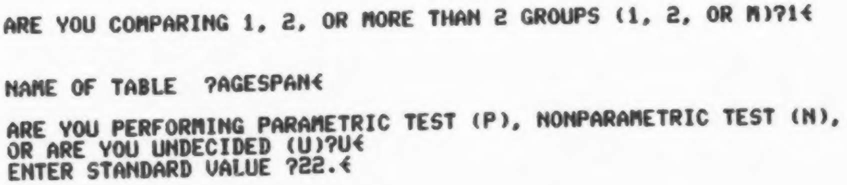


AGESPAN: AGE AT DEATH 25R X 1C

	1 YEARS		1 YEARS
1	17.2	14	24.2
2	18.	15	24.6
3	18.7	16	25.8
4	19.8	17	26.
5	20.3	18	26.3
6	20.9	19	27.2
7	21.	20	27.6
8	21.7	21	28.1
9	22.3	22	28.6
10	22.6	23	29.3
11	23.1	24	30.1
12	23.4	25	35.1
13	23.8		

Figure 3-2

We use the COMPARE dialogue to make this test (Figure 3-3a-b). When we specify that we are UNDECIDED whether the data come from a normal population, the test for normality is automatically performed. Since there is no evidence that the population is nonnormal the parametric test is performed. COMPARE asks us to supply the standard value; we indicate 22 years. COMPARE displays the results of the one sample t-test in a table. The calculated t-value is 2.623 with n-1=24 degrees of freedom. Since this exceeds the critical t-value at the 0.05 level, the null hypothesis is rejected. That is, this sample gives evidence that the population mean life span for this group of horses is not 22 years.



```
ARE YOU COMPARING 1, 2, OR MORE THAN 2 GROUPS (1, 2, OR M)?1←

NAME OF TABLE  ?AGESPAN←

ARE YOU PERFORMING PARAMETRIC TEST (P), NONPARAMETRIC TEST (N),
OR ARE YOU UNDECIDED (U)?U←
ENTER STANDARD VALUE ?22.←
```

Figure 3-3a

```
T-TEST OF MEAN VS STANDARD

AGESPAN: AGE AT DEATH  (NOTES)
+------------+---+------+--------+
|            |N  |MEAN  |STDEV   |
+------------+---+------+--------+
|1. YEARS    |25.|24.228|4.247208|
+------------+---+------+--------+
|2. STD VALUE|   |22.   |        |
+------------+---+------+--------+

POPULATION COLUMN 'YEARS' ARE NORMAL AT P>0.05 BY W-STATISTIC.

NULL HYPOTHESIS  H0:   THE POPULATION MEAN AND STANDARD ARE EQUAL.

T = 2.623   DF = 24   P = .0149
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05.
```

Figure 3-3b

In any hypothesis test involving the t-distribution, when the calculated t-value is less than the critical t-value the null hypothesis is not rejected. Conversely, when the critical value is met or exceeded, there is sufficient evidence to reject the null hypothesis. It is important to recognize that acceptance of the null hypothesis is not equivalent to stating that it is true, merely that there was insufficient evidence to reject it.

B. The t-test for paired data

In comparing two samples, researchers often try to match items or subjects so that they are paired. One item in each pair is then randomly assigned to one experimental condition and the other item to a second condition. For example, two subjects with similar characteristics, such as age and sex, might be paired and assigned to different treatments. To obtain even greater precision, a single group of subjects may be tested under two conditions, and the responses compared.

Pairing has the advantage that the differences among subjects are spread evenly over the two treatments. If the two treatments are truly different, then the difference will not be obscured by differences among the subjects. Conversely, if the treatments are effectively the same, pairing reduces the chance that differences between subjects will randomly result in an apparent difference between treatments.

By considering the set of differences between pairs of data as a single sample, you can then test whether the mean of these differences is significantly different from 0. The paired t-test is thus equivalent to testing a single sample to a standard of value equal to 0 (as discussed in Section 3.4A).

To compare paired samples, you would again use the COMPARE command. In this case, you specify that there are two samples to be compared. You can use data from one or two tables, organized by rows or columns. You respond "YES" to the question "ARE THE DATA PAIRED?". Figure 3-4 below summarizes the results for the paired-t analysis (Zar, p. 122).

```
T-TEST ON PAIRED DATA

DEER: LENGTH OF LIMB (CM)  (NOTES)
+---------------+----+-----+--------+
|               |N   |MEAN |STDEV   |
+---------------+----+-----+--------+
|1. FORE        |10. |144.7|3.400974|
|2. HIND        |10. |141.4|4.033201|
|3. DIFFERENCE  |10. |3.3  |3.056868|
+---------------+----+-----+--------+

POPULATION DIFFERENCES ARE NORMAL AT P>0.05 BY W-STATISTIC.

NULL HYPOTHESIS  H0: THE POPULATION MEANS ARE EQUAL.

T = 3.414   DF = 9   P = .0077
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.
```

Figure 3-4

In a paired t-test, each observation in one group must have a paired observation in the second group. If the analysis is performed on samples that have missing observations, the unmatched data in question will be excluded; this will be noted below the results table.

The paired sample t-test makes no assumptions about the normality of the two samples individually; rather, it assumes that the differences come from a normally distributed population. The paired-sample layout is thus a very powerful one, because it is not explicitly affected by the variability within each sample.

C. Unpaired t-test

In many situations an experimental design consisting of paired data is not feasible. In these cases the unpaired or two-sample t-test is employed. This test takes one of two forms, depending on whether or not the variances of the two populations from which the samples were drawn are considered to be different. When they are not different, the equal variance t-test is used. Otherwise, the unequal variance t-test is used.

The first step in an unpaired t-test is thus to test for the equality of variances. This is done with the F-test. The F-statistic is defined as

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 is greater than s_2^2 (otherwise the order is simply reversed). The degrees of freedom are (n_1-1) and (n_2-1). The null hypothesis is that the variances are not different. If the calculated F value exceeds the critical F-value, the null hypothesis is rejected; if not, it is accepted.

When it is concluded that the variances are equal, the appropriate t-statistic with degrees of freedom = $n_1 + n_2 - 2$, is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2 (1/n_1 + 1/n_2)}}$$

where s_p^2, the pooled estimate of the population variance σ^2, is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

When the variances are not equal, then

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_2 + s_2^2/n_2}}$$

$$DF = \text{CFIXED}\left[\frac{(s_1^2/n_1 \quad + \quad s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}\right]$$

When the variances are not equal it is not clear how to obtain a single pooled value for the variance of the differences in the means. When one mean is not known with the same accuracy as the other, it is not obvious what would be the "best" estimate for the variance of the difference of the two means. This is the Behrens-Fisher problem that is controversial to this day. The above test is merely one of several ways of comparing the two means when the population variances are not equal. It is usually better to transform the data (e.g., with logarithms, square roots, squares) so that the variances of the samples are almost the same and use the equal variance t-test than to use the unequal variance t-test.

In Figure 3-5, we use COMPARE to perform the unpaired t-test on the pH levels of two ponds (Zar, p. 104). COMPARE automatically performs the F-test and reports its results. If the null hypothesis (variances are equal) is accepted, the equal variance procedure is performed. If the null hypothesis is rejected, you are asked if you wish to use the unequal variance test. If you respond "NO" to this question, the equal variance test will be used; otherwise the unequal variance test will be performed. The results of the unpaired t-test, equal variance case, are shown in Figure 3-5 below.

```
T-TEST ON UNPAIRED DATA (EQUAL VARIANCE)

MOTH  (NOTES)
|           || N   | MEAN     | STDEV    |
|1. TRAP 1  || 11. | 34.45455 | 4.676829 |
|2. TRAP 2  || 8.  | 57.25    | 3.918819 |

POPULATION COLUMN 'TRAP 1' ARE NORMAL AT P>0.05 BY W-STATISTIC.
POPULATION COLUMN 'TRAP 2' ARE NORMAL AT P>0.05 BY W-STATISTIC.

F = 1.424   DF(N/D) = 10/7   P = .328

NULL HYPOTHESIS  H0: THE POPULATION MEANS ARE EQUAL.

T = 2.716   DF = 17   P = .0147
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05.
```

Figure 3-5

3.5 Nonparametric tests

For many nonparametric procedures the actual measured values are not employed. Instead, the measurements are ranked in order and it is the distribution of the ranks that forms the basis for the test. This operation makes the actual values of the data unimportant.

A. One-sample test

The Wilcoxon signed rank test is used to test inferences about the central tendency of the population from which the sample was drawn. This is done by comparing the sample to a known standard or hypothetical value. Although this test makes no assumptions about the distribution of the data (i.e., normal or otherwise), it does assume that the population is continuous and symmetric about the central measure. Thus, for the test to be valid, you should first look at a histogram of the data. If it is highly skewed you should try to transform the data.

The first step in the test is to calculate the differences between the sample data and the standard value. Disregarding the sign, these differences are then ranked in order. Finally, the ranks are signed according to whether the differences are positive or negative.

Let T be the sum of the ranks with a positive sign and T' be the absolute value of the sum of the ranks with negative sign. It may be noted that T and T' are related by the following equation:

$$T' = \frac{n(n+1)}{2} - T$$

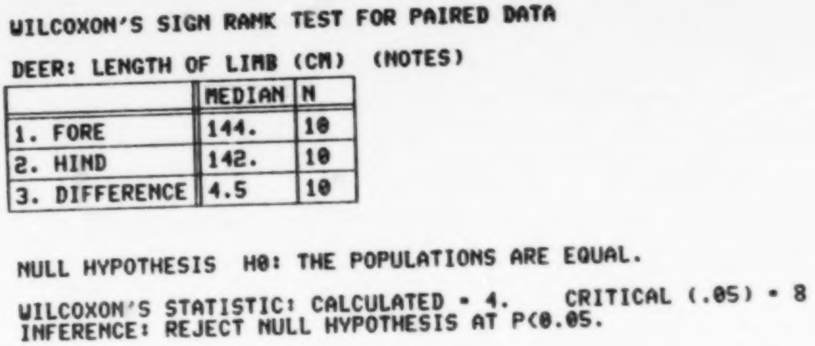
where n is the size of the sample. If either T or T' is less than or equal to the critical value for the test, the null hypothesis is rejected. Tied observations are assigned the mean of the ranks they would have received if they were not tied. Zero differences are ignored. Notice that in contrast to the t-tests discussed in the previous section, small values of the calculated T or T' are significant. That is, when T and T' are very different there is evidence that the data tend to be to one side or the other of the standard value.

For large values of n (> 100) the distribution of T approximates a normal distribution, and the following Z-statistic may be used:

$$Z = \frac{\left| T - n(n+1)/4 \right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

B. Paired-sample test

The method used for the paired-sample case is another form of the Wilcoxon signed rank test. It is identical to the one-sample test discussed in Section 3.5A, except that the differences between pairs of data values (as opposed to the difference between the values of the data and the standard) are ranked. This time the assumption about the distribution is that the sample of differences is symmetric. An example comparing the length of forelimbs and hindlimbs in deer is shown in Figure 3-6 (Zar, p. 125).



```
WILCOXON'S SIGN RAMK TEST FOR PAIRED DATA

DEER: LENGTH OF LIMB (CM)  (NOTES)
+---------------+--------+----+
|               | MEDIAN | N  |
+===============+========+====+
| 1. FORE       | 144.   | 10 |
| 2. HIND       | 142.   | 10 |
| 3. DIFFERENCE | 4.5    | 10 |
+---------------+--------+----+

NULL HYPOTHESIS  H0: THE POPULATIONS ARE EQUAL.

WILCOXON'S STATISTIC: CALCULATED = 4.    CRITICAL (.05) = 8
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05.
```

Figure 3-6

C. Unpaired-sample test

To analyze samples in the nonparametric case when the data are not paired, PROPHET uses the form of the Wilcoxon rank sum test known as the Mann-Whitney test. A preliminary assumption is that the dispersions of the two samples are not different, and this is automatically tested using the Ansari-Bradley test. When you are performing the two-sample test and the dispersions are indicated to be unequal, COMPARE gives you the option of not proceeding further with the analysis. In this case transformation of the data is recommended.

In the Wilcoxon rank sum test, the two samples are ordered together and the sum of the ranks for each sample is found. If one sample has data which are generally larger than the other, it will also have a larger rank sum than the other. The test statistics are

$$U = n_1 n_2 + n_1(n_1 + 1)/2 - R_1$$

$$U' = n_1 n_2 - U$$

R_1 is the sum of the ranks of observations in group 1.

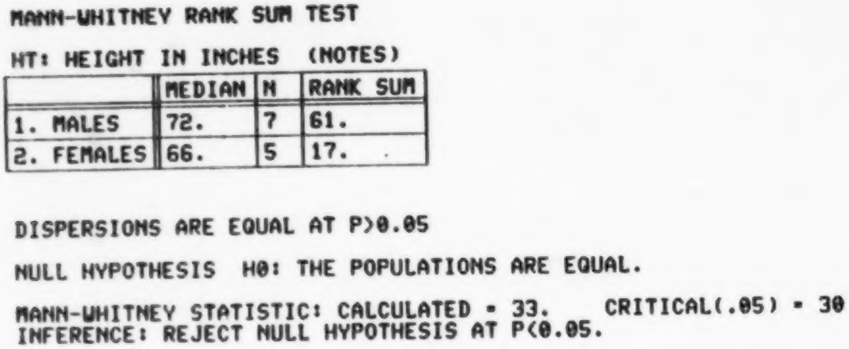
The larger of U and U' is then compared to the critical value, and if the critical value is exceeded, the null hypothesis is rejected.

When the smaller sample is of size >=20, and the larger is of size >=40, the normal approximation is used:

$$z = \frac{|U - n_1 n_2/2|}{\sqrt{\frac{n_1 n_2}{N^2 - N}\left(\frac{N^3 - N - \Sigma(t_j^3 - t_j)}{12}\right)}}$$

where t_j is the number of ties in a group of tied values and $N = n_1 + n_2$.

An example of the Mann-Whitney test is shown in Figure 3-7.

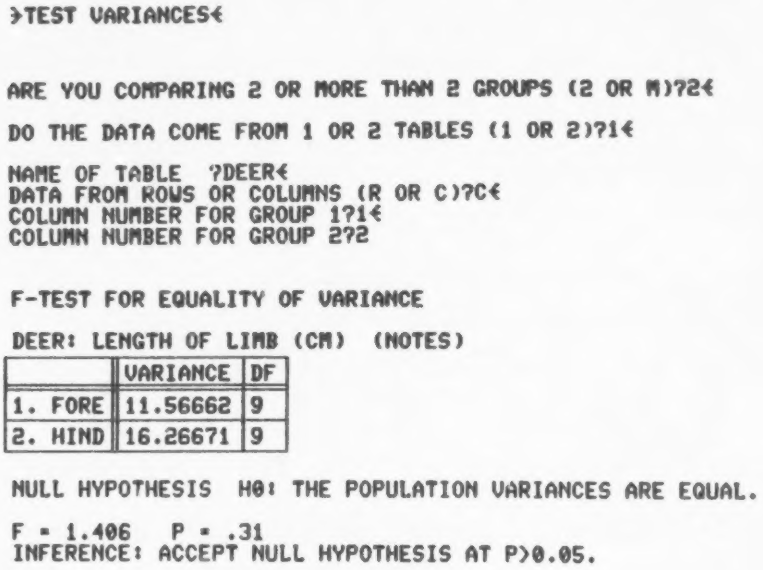


```
MANN-WHITNEY RANK SUM TEST

HT: HEIGHT IN INCHES  (NOTES)
+----------+------+-+---------+
|          |MEDIAN|N|RANK SUM |
+----------+------+-+---------+
|1. MALES  |72.   |7|61.      |
|2. FEMALES|66.   |5|17.      |
+----------+------+-+---------+

DISPERSIONS ARE EQUAL AT P>0.05

NULL HYPOTHESIS  H0: THE POPULATIONS ARE EQUAL.

MANN-WHITNEY STATISTIC: CALCULATED = 33.    CRITICAL(.05) = 30
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05.
```

Figure 3-7

3.6 PROPHET Usage: TEST VARIANCES

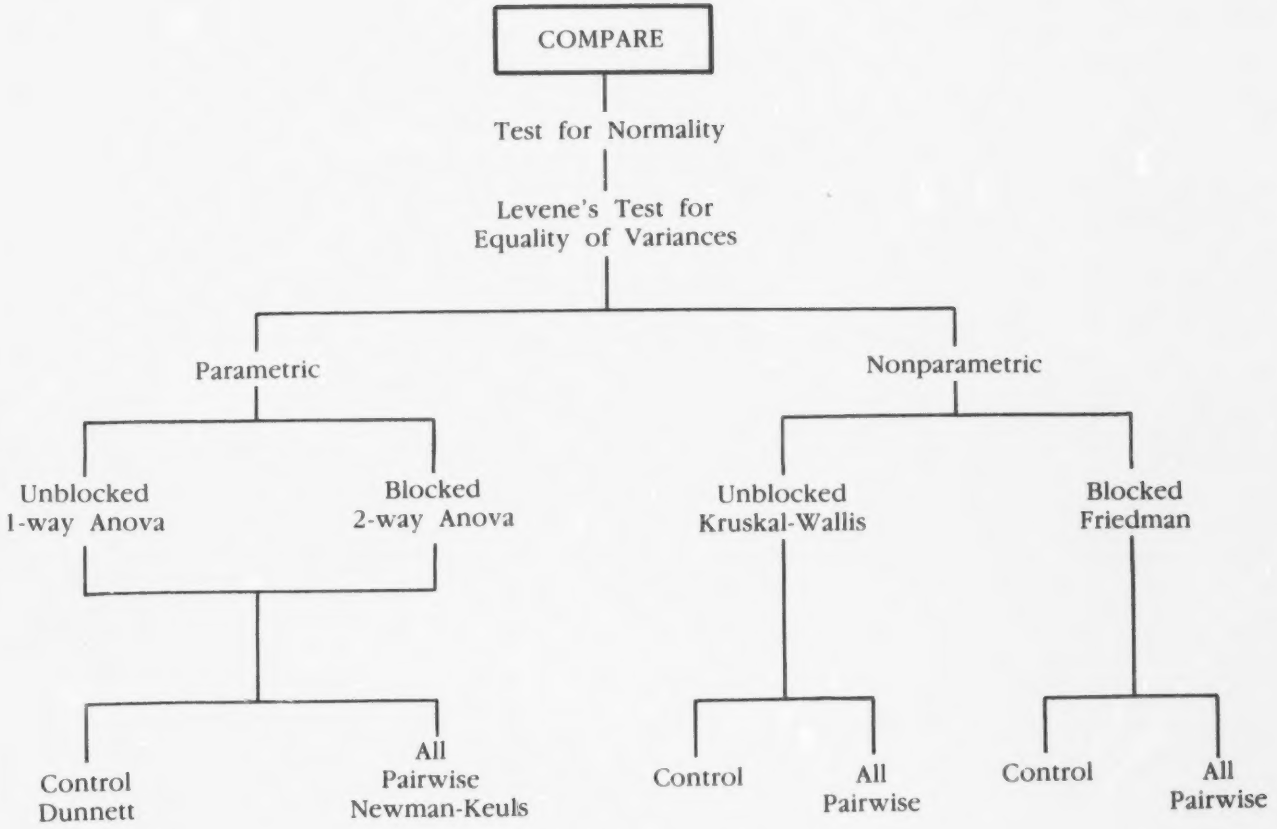
As we discussed above in connection with the two sample t-test, COMPARE uses the F-test to test the null hypothesis that two variances are not significantly different. The F-test can also be performed in PROPHET with the TEST VARIANCES command. In TEST VARIANCES, you are asked for the name of the data table, whether the data are organized by rows or columns, and so on, just as you are in the COMPARE dialogue. The results of the test are shown in Figure 3-8.



```
>TEST VARIANCES<


ARE YOU COMPARING 2 OR MORE THAN 2 GROUPS (2 OR M)?2<

DO THE DATA COME FROM 1 OR 2 TABLES (1 OR 2)?1<

NAME OF TABLE  ?DEER<
DATA FROM ROWS OR COLUMNS (R OR C)?C<
COLUMN NUMBER FOR GROUP 1?1<
COLUMN NUMBER FOR GROUP 2?2


F-TEST FOR EQUALITY OF VARIANCE

DEER: LENGTH OF LIMB (CM)  (NOTES)
+---------++----------+--+
|         || VARIANCE |DF|
+---------++----------+--+
| 1. FORE || 11.56662 |9 |
+---------++----------+--+
| 2. HIND || 16.26671 |9 |
+---------++----------+--+

NULL HYPOTHESIS  H0: THE POPULATION VARIANCES ARE EQUAL.

F = 1.406   P = .31
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.
```

Figure 3-8

4. COMPARE: II


COMPARE
Test for Normality
Levene's Test for Equality of Variances
Parametric
Nonparametric
Unblocked 1-way Anova
Blocked 2-way Anova
Unblocked Kruskal-Wallis
Blocked Friedman
Control Dunnett
All Pairwise Newman-Keuls
Control
All Pairwise
Control
All Pairwise

Section 4

MULTISAMPLE HYPOTHESIS TESTING

4.1 Introduction

The methods of hypothesis testing discussed in Section 3 are appropriate for comparisons of two samples, or of one sample versus the value of a known standard. Different techniques are required, however, when an experimental design consists of more than two samples. Multiple samples may not be compared by simply repeating a two sample test many times to accommodate all pairings of groups.

Consider, for example, when three samples are drawn from the same population. The probability of rejecting the null hypothesis is greater for the smallest and largest means than for any other pair. If next you were to draw ten samples from the same population, then by chance alone you would expect the spread between the smallest and the largest samples to be larger still.

Hence, in performing a series of two-sample t-tests the likelihood of finding at least one significantly different pair by chance increases with the number of comparisons being made. For example, suppose three samples are compared with three t-tests (A vs. B, A vs. C, and B vs. C). If the significance level is set at 0.05 for each comparison, the overall probability of committing a Type I error is actually 13%, not the intended 5%. This error rate continues to rise dramatically as the number of pairs (i.e., number of repeated tests) is increased. Thus, in multiple sample comparisons, a different approach is needed to appropriately consider all the data simultaneously.

The first step is to test the null hypothesis that the samples were drawn from populations that have the same mean or central tendency. This is done by analysis of variance techniques that are extensions of the two-sample t-tests or by analogous extensions of nonparametric two-sample tests. The nonparametric tests are very similar to analysis of variance except that ranks are used instead of the actual data values. In either case (i.e., parametric or nonparametric), if there is sufficient evidence to reject the null hypothesis then further analysis is needed to reveal specifically among which populations the differences occur. This is done by comparing two selected samples to determine if they are different. The difference between these multiple comparison procedures and, for example, the repeated use of the t-test as described above, is that the error rate is maintained at a fixed level α. Taking into account the number of comparisons made about pairs of samples simultaneously, the probability of concluding somewhere among the comparisons that two populations are not equal, when they in fact are, is α.

4.2 PROPHET Usage: COMPARE

The COMPARE command, introduced in Section 3, also performs parametric analysis of variance, analogous nonparametric multisample tests, and the appropriate multiple comparison procedures. Several questions are asked in the COMPARE dialogue to determine the organization of your data and which tests are appropriate to perform: in other words, which path of the roadmap facing page 4-1 the procedure will take. Figure 4-1 shows the initial dialogue of the COMPARE command. The meaning of each question is explained below.

```
>COMPARE<
ARE YOU REPEATING THE SAME TEST PROCEDURE ON SEVERAL SETS OF DATA?N<

ARE YOU COMPARING 1, 2, OR MORE THAN 2 GROUPS (1, 2, OR M)?M<
NAME OF TABLE?IQ<
DATA FROM ROWS OR COLUMNS (R OR C)?C<
DO YOU WISH TO PERFORM THE ANALYSIS ON THE ENTIRE TABLE ?N<
DO YOU WISH TO EXCLUDE ROWS (R) COLUMNS (C) OR BOTH (B) ?B<
ENTER COLUMN  NUMBERS TO BE EXCLUDED SEPARATED BY SPACES
OR COMMAS?2<
ENTER ROWS TO BE EXCLUDED ?3 5 6<

ARE YOU PERFORMING PARAMETRIC TEST (P), NONPARAMETRIC TEST (N),
OR ARE YOU UNDECIDED (U)?P<

ARE THE DATA BLOCKED?N<
IS THERE A CONTROL GROUP?Y<
ENTER COLUMN NUMBER FOR THE CONTROL ?1<
```

Figure 4-1

ARE YOU REPEATING THE SAME TEST PROCEDURE ON SEVERAL SETS OF DATA?

The meaning of this question is explained in Section 3.3. If you have many similar sets of data, all of which are to be analyzed in the same way, you enter "YES" followed by the total number of data sets. Otherwise you enter "NO".

ARE YOU COMPARING 1, 2, OR MORE THAN 2 GROUPS (1, 2 OR M)?

Since the procedures described in this section are used only for comparing more than two samples, you respond "M".

NAME OF TABLE?

You enter the table name.

DATA FROM ROWS OR COLUMNS (R OR C)?

If your samples are contained in the columns of the table, you enter "C"; if they are organized by rows, enter "R".

DO YOU WISH TO PERFORM THE ANALYSIS ON THE ENTIRE TABLE?

If you wish to include all of the data in the table, respond "YES". Occasionally you may wish to exclude a certain portion of the table. If you respond "NO" you are asked if you wish to exclude rows, columns or both. You respond with "R", "C", or "B". You will then be prompted for the rows/columns to be excluded; type the numbers separated by spaces or commas. Note: You do not need to specifically exclude columns that contain nonnumeric data, or rows or columns which contain statistical measures added by STATS or MEASURES commands. This will be done automatically.

ARE YOU PERFORMING PARAMETRIC TEST (P), NONPARAMETRIC TEST (N), OR ARE YOU UNDECIDED (U)?

At the top of the roadmap facing page 4-1 a test for normality of the data is indicated, to determine if parametric or nonparametric procedures are appropriate. If you wish to have COMPARE test for normality you respond "UNDECIDED (U)", in which case the appropriate test (D'Agostino or Wilk-Shapiro, depending on sample size) will be performed, and the results presented. When the results indicate that the data are nonnormal, you have the option of continuing with parametric tests, performing nonparametric tests, or leaving the program. As you can see on the roadmap, if you assume normality you will be performing the parametric tests on the lefthand side; otherwise you will be performing the nonparametric tests on the righthand side.

ARE THE DATA BLOCKED?

This question pertains to the design of the experiment from which the data were drawn; it is simply the extension of the question "ARE THE DATA PAIRED?" to the case of more than two samples. For example, if the data are organized by columns and each row represents a given subject who receives all of the treatments being studied, then the data are blocked. If on the other hand the data are randomly assigned to each sample, so that no relationship exists, say, between the first observation in sample 1 and the first observation in sample 2, then the data are unblocked. The unblocked case is the extension of the unpaired two-sample layout. Thus you respond "YES" or "NO" to the above question, depending on the experimental design.

IS THERE A CONTROL GROUP?

To determine specifically which samples are different there are two possible approaches, depending on the purpose of the experiment. The first is to perform all pairwise comparisons among samples. The second is to compare a single specified sample, termed the control, with all other samples. If you respond "YES" to this question you then specify the column number (or row number) for the control sample.

Levene's Test

After you complete the dialogue, COMPARE will perform Levene's test for homogeneity of variance of the data. Technically, this is a parametric test, but it is also very robust and may be used for nonnormal data. Data which show very strong evidence against the null hypothesis that the population variances are equal, should be re-examined for possible outliers or nonadditivity of effects. But it is worth noting that the analysis of variance is robust enough to be used even with considerable heteroscedasticity as long as the sample sizes of the different groups are nearly equal.

Parametric Procedures

On the parametric (normal) side of the roadmap, either a one-way analysis of variance for unblocked data or a two-way (randomized block) analysis of variance is performed. The analysis of variance table is displayed, along with the usual inference statement regarding acceptance or rejection of the null hypothesis being tested.

In the one-way analysis of variance the sample sizes need not be equal (although it is good statistical practice to make them so, if possible). The two-way analysis of variance, however, requires that the sample sizes be equal (i.e., that the blocks be complete). Thus, when missing values are encountered they will automatically be replaced with values calculated to minimize the error mean square.

When the null hypothesis is accepted, you are advised not to continue with the multiple comparison tests, but you are given the option of doing so if you wish. If you indicated that there is not a control group, then COMPARE performs all comparisons of two samples using the Newman-Keuls test. If you specified a control, then that sample will be compared to all others using Dunnett's test. Both of these procedures use the error mean square found in the overall analysis of variance test.

Nonparametric Procedures

On the nonparametric side, the Kruskal-Wallis test will be performed when the data are unblocked; the Friedman test will be performed when the data are blocked. Both of these tests can be extended to do either all pairwise comparisons or comparisons with a control. If the null hypothesis in the overall test is accepted, you are given the option of performing the multiple comparison tests; otherwise, they are performed automatically. These procedures parallel the methodology of the Newman-Keuls and Dunnett's tests; however, the sample rank sums are used instead of the means.

The number of observations for each sample need not be the same for the Kruskal-Wallis test. However, the sample sizes must be equal for the Friedman test, and blocks in which missing data (i.e., empties) are found will automatically be excluded. This information will be reported along with the results of the test.

Examples of each of these tests and further discussion of the COMPARE output are described in the following sections.

4.3 Principles of Analysis of Variance

This section provides a brief introduction to the major principles and assumptions of the analysis of variance. While we discuss specifically the one-way analysis of variance of unblocked data, the underlying concepts are applicable to higher order designs as well.

Analysis of variance is a technique used to determine if the mean values of sets of observations (samples) derived under different experimental conditions could all have been drawn from the same population. These experimental conditions are controlled by the investigator and are termed fixed effects. In this model, therefore, the jth observation in the ith sample, can be expressed as

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Here μ is the true mean of the population, α_i is the effect of the ith treatment (or experimental condition) and ε_{ij} is a normally distributed random error variable with mean equal to zero and variance equal to σ_i^2, the variance of the population. Thus the analysis of variance tests whether or not all the effects, α_i, are equal to zero. When they are different, the null hypothesis that the samples were all drawn from the same population is rejected.

By way of illustration, consider the case of three samples. Figure 4-2 shows two possible distributions of three samples. In the upper, the three samples are drawn from three different populations with means μ_1, μ_2, μ_3 and variances σ_1^2, σ_2^2, σ_3^2; in the other case the samples are drawn from the same population with mean μ and variance σ^2.

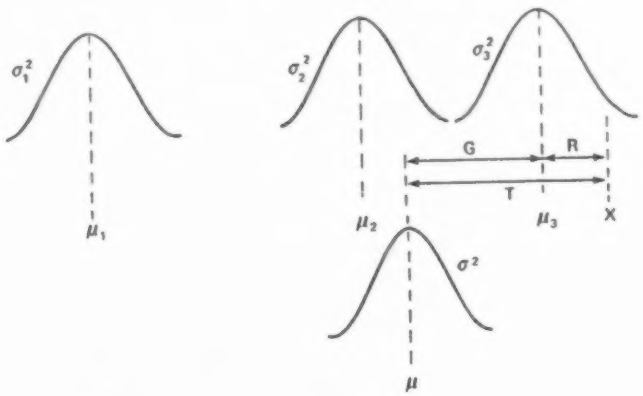


Figure 4-2

Let X_{ij} correspond to observation j on sample i and n_i be the number of observations in sample i. The estimate of the variance for sample 1 is:

$$s_1^2 = \frac{\sum_{j=1}^{n_1} (X_{1j} - \overline{X}_{1.})^2}{n_1 - 1}$$

where $\overline{X}_{1.}$ is the mean of the observations in sample 1. Similarly, for samples 2 and 3 the estimate of the mean and variance can be computed. If $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ then you are justified in pooling the three estimates of variance to obtain an overall estimate of the variance σ^2, in exactly the same way as is done in the two-sample t-test with equal variances. Thus the best estimate for variance, called the error mean square, or residual mean square, is:

$$s_I^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})^2}{N - k}$$

where N is the total number of observations and k is the number of samples. This is an average squared distance of each observation from its sample mean. The degrees of freedom for the estimate is N-k.

But when the null hypothesis is true, i.e., $\mu_1 = \mu_2 = \mu_3$, there is another way to estimate the variance σ^2. Recall that the variance of the sample mean $\overline{X}_{1.}$ is

$$\sigma^2_{\bar{x}_1} = \sigma^2/n_1$$

Rewriting this formula, we get

$$\sigma^2 = n_1\sigma^2_{\bar{x}_1}$$

Thus an estimate of the variance of the sample mean provides an estimate of the population variance (similarly for samples 2 and 3). The second estimate of s_p^2 is obtained by pooling these estimates to obtain the groups mean square. Here the variability of the individual means about the overall mean is described.

$$s_{II}^2 = \frac{\sum_{i=1}^{k} n_i(\overline{X}_{i.} - \overline{X}_{..})^2}{k - 1}$$

Under the assumptions that the population variances are equal and the null hypothesis is true, these two estimates are estimates of the overall variance and you should expect their ratio to be close to unity. Conversely, if the null hypothesis is not true then the means of each sample will vary significantly both from each other and from the overall mean. However, s_I^2 is not influenced by the differences in the means since the deviations are measured within each sample. Thus you should expect s_{II}^2 to be much greater than s_I^2. The ratio of s_{II}^2 and s_I^2 is equal to F. If F is large enough you would conclude that the null hypothesis was not true. If the data come from normal populations, then under the null hypothesis this ratio follows the F distribution (similar to the F-test used to compare two variances in Section 3).

There is an interesting relationship between these two estimates of variance:

$$\Sigma\Sigma\,(X_{ij}-\overline{X}_{..})^2 \;=\; \Sigma\Sigma\,(X_{ij}-\overline{X}_{i.})^2 \;+\; \Sigma\Sigma\,(\overline{X}_{i.}-\overline{X}_{..})^2$$

$$\text{Total SS} \;=\; \text{Residual SS} \;+\; \text{Group SS}$$

Notice that

$$s_{I}^2 = \text{Residual SS}/(N-k)$$

and

$$s_{II}^2 = \text{Group SS}/(k-1)$$

N-k and k-1 are called the residual degrees of freedom and the group degrees of freedom respectively. The total degrees of freedom is N-1.

From this you can see that the total variability of the data about the overall mean $\overline{X}_{..}$ can be divided into two parts, the residual sum of squares and the group sum of squares. In Figure 4-2 the different components for the observation X are marked T, R, and G corresponding to total, residual and group deviations respectively.

In more complicated designs the total sum of squares can be divided into more components allowing for better estimates of the residual and for tests of additional hypotheses. This point will be explained further in the discussion of two-way analysis of variance below and in Section 7.

In general, the test procedures for the analysis of variance consist of two parts. First, the different sources of variation are calculated, and their degrees of freedom and significance levels are assessed. The null hypothesis for the analysis of variance is that the samples were drawn from the same population. This is equivalent to saying that all treatment effects are identical. Rejection of the null hypothesis, while indicating that differences do exist, does not individually identify them. When the null hypothesis is rejected, the second stage of the analysis is therefore to perform a multiple comparison procedure, to determine precisely where the differences occur.

4.4 Assumptions

In all the tests (parametric and nonparametric) used to compare the measures of central tendency of several samples, the following assumptions are made:

- Effects are additive.
- The errors are mutually independent.
- The errors come from the same continuous population.

In addition, the parametric tests based on the analysis of variance have a further restriction on the third assumption: that the errors follow a normal distribution.

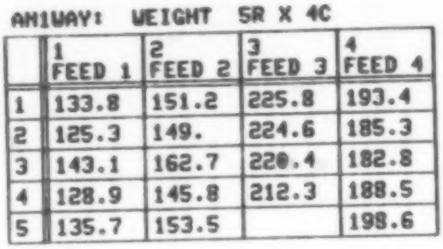
The first assumption is really a restatement of the model that the observations are the sum and not the product of the population mean and the specified effect. One way of detecting nonadditive data is to check whether the coefficients of variation (see Section 1) for the different samples are approximately the same. If they are, then a logarithmic transformation on the data will correct this problem. (See Snedecor & Cochran, p. 330). If this is not corrected, two problems arise. First, the real nature of the data is not discovered: that is, that the effects are proportional. Second, the analysis will be insensitive to differences because the error is inflated.

The second assumption is that the error terms associated with each observation are independent. Lack of independence of the error terms may occur because of inadequate randomization of the observations or a significant trend in the observations. For example, a measuring device may consistently overestimate the true value by a sequentially increasing amount. This can be corrected by randomly assigning the order of observations of all the treatments; the mean and variance of the errors should be the same for each group. In general, conservative randomization and careful experimental design ensure that the errors will be independent.

The final assumption is that the distribution of the errors is the same for all the treatments. If the errors for one sample are greater in one treatment than another then the difference may erroneously be attributed to a treatment effect. For example, the effects of an experiment relating diet to weight change in rats would clearly be spurious if the five lightest rats were assigned to one diet, the five heaviest rats to another. Again, randomization would reduce the likelihood of the errors having different central measures for different treatments. In the same way that the dispersions should be equal in two-sample testing, the dispersions should be homogeneous for multisample tests. Levene's test is available in PROPHET for checking for badly nonhomogeneous dispersions. Making histograms of the data for each treatment is also a good way to check whether the shapes of the samples' distributions are the same.

4.5 One-way Analysis of Variance

The simplest form of the analysis of variance is the completely randomized design. It examines the effects of a single factor on a single variable. For example, in studying the effects of different feed types on the weight of pigs, the single factor is feed type, and the single variable is weight gain. Each type of feed is known as a level of the factor. The pigs are assigned randomly to each feed type to help eliminate any bias: hence the name completely randomized design. The data for this example (Zar, p.134) are shown in Figure 4-3 and the resultant output is shown in Figure 4-4.



AN1WAY: WEIGHT 5R X 4C

	1 FEED 1	2 FEED 2	3 FEED 3	4 FEED 4
1	133.8	151.2	225.8	193.4
2	125.3	149.	224.6	185.3
3	143.1	162.7	220.4	182.8
4	128.9	145.8	212.3	188.5
5	135.7	153.5		198.6

Figure 4-3

As discussed above, the purpose of the analysis of variance is to separate into several components the different contributions to the total variability of the data. Each component is calculated as a "sum of squares" quantity (SS) which, when divided by its degrees of freedom, is analogous to the sample variance. The first component in this example is the individual variability within each group of pigs (i.e., the individual variability among those that receive Feed 1, those that receive Feed 2, etc.). Since presumably the individual pigs in each group respond similarly to the diet, this term is expected to be small, and to comprise merely experimental or random errors. Thus, for each group the sum of squared deviations about the group mean are calculated; these quantities are then summed to give the resultant source of variation, the residual SS. If this quantity is divided by its degrees of freedom (N-k, k=number of groups), the resultant residual mean square is the best estimate of the variance common to all the different populations. The calculation of these quantities is shown in Table 4-1.

```
ANALYSIS OF VARIANCE ON AN1WAY
WEIGHT  (NOTES)
+---------------+----------+----+----------+----------+------------+
|SOURCE         |SUM OF    |DF  |MEAN      |F         |SIGNIFICANCE|
|               |SQUARES   |    |SQUARE    |STATISTIC |LEVEL       |
+---------------+----------+----+----------+----------+------------+
|1. COLUMN      |20461.41  |3   |6820.471  |164.376   |.0001       |
|2. RESIDUAL    |622.3984  |15  |41.49323  |          |            |
|3. TOTAL       |21083.81  |18  |          |          |            |
+---------------+----------+----+----------+----------+------------+

LEVENE'S STATISTIC = .036.  VARIANCES ARE HOMOGENEOUS AT P>0.05.

COLUMN 'FEED 1' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'FEED 2' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'FEED 3' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'FEED 4' IS NORMAL AT P>0.05 BY W-STATISTIC.


NULL HYPOTHESIS  H0: THE POPULATION MEANS ARE EQUAL.

INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.
```

Figure 4-4

The second component of variation is a measure of the variability of the individual sample means about the total mean. This term is called the Groups SS. In Figure 4-4, PROPHET denotes it the Column SS, to indicate the fact that the data (samples) were organized by columns for this example. If the feeds in this example have different effects, one would expect this component to be large. Dividing this term by its degrees of freedom (k-1) results in the column (group) "mean square".

Finally, the results table shows the total sum of squares, which is equal to the variability of all the data about the total mean. Note the fundamental relationship between the three sum of squares terms:

$$\text{Total SS} = \text{Column SS} + \text{Residual SS}$$

The null hypothesis for this example is that the population means of the four samples are not different. This single hypothesis is tested using the F-statistic:

$$F = \frac{\text{Column Mean Square}}{\text{Residual Mean Square}}$$

with k-1 and N-k degrees of freedom. The significance level associated with the F-statistic is shown as well in the results table (Figure 4-4). The usual inference statement about the null hypothesis based on this probability is reported below the table; in this case the null hypothesis is rejected.

It should also be noted that the data in each sample were tested for normality using the Wilk-Shapiro (W-S) statistic. (Had the data consisted of more than 50 obervations, the normality test would have been based on the D'Agostino (D) statistic.)

It is worth noting that the one-way, completely randomized analysis of variance is an extension of the two-sample t-test. In fact, for the case of two groups, the two test procedures are equivalent.

TABLE 4-1. ONE-WAY ANALYSIS OF VARIANCE.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Value
Group	$\sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$	k − 1	(Group SS)/(k − 1)	(Group MS)/(Residual MS)
Residual	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$	N − k	(Residual SS)/(N − k)	
Total	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{\cdot\cdot})^2$	N − 1		

4.6 Two-way Analysis of Variance

Certain experimental designs allow an additional component of overall data variability to be identified. For example, consider a study designed to assess the chronotropic properties of three different drugs. One experimental design would be to assign one of the drugs at random to each of twelve individuals, such that four patients received each drug. The results of these twelve observations would be analysed by the one-way test described in the previous section. It is known, however, that the response to any drug will vary from individual to individual. This inherent variability is not of interest to the investigators and in fact may tend to obscure the effects of the drugs being tested. Thus, an alternative approach would be to employ only four test subjects, giving them all each of the drugs in a random order. In this case, the resulting data are said to be "blocked", and are appropriately analysed by the two-way, completely randomized block analysis of variance. Each subject constitutes a "block", and an additional source of variation may be attributed to the variability within these blocks. The model for this analysis is

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where α_i represents the "effect" or "treatment", and β_j is the block term. (ε_{ij} is the error term.)

Just as the one-way, completely randomized analysis of variance is an extension of the two-sample t-test, so is the two-way randomized block analysis of variance an extension of the paired t-test.

An example of the data for a randomized block design is shown in Figure 4-5.

Note that to perform the two-way analysis of variance you respond "YES" to the question "ARE THE DATA BLOCKED?". The results for the two-way analysis are shown in Figure 4-6. The calculations of the various quantities in the results table of the two-way analysis of variance are shown in Table 4-2. Note that PROPHET will always put the terms referring to groups in the first row of the analysis of variance table and the block terms in the second row. In this example the data were organized by columns and the label "ROW" in Figure 4-6 denotes the block structure.

The null hypothesis that the true population means are not different is tested by evaluating the F-statistic as the ratio of the group (column) mean square and the residual mean square.

best copy available

DIETS: WEIGHT GAIN (GMS) 11R X 4C

0 ROWNAME	1 DIET 1	2 DIET 2	3 DIET 3	4 DIET 4
1. BLOCK 1	1.5	2.7	2.1	1.3
2. BLOCK 2	1.4	2.9	2.2	1.
3. BLOCK 3	1.4	2.1	2.4	1.1
4. BLOCK 4	1.2	3.	2.	1.3
5. BLOCK 5	1.4	3.3	2.5	1.5
6. -N-	5.	5.	5.	5.
7. -SUM-	6.9	14.	11.2	6.2
8. -MEAN-	1.38	2.8	2.24	1.24
9. -VARIANCE-	.01199999	.1999999	.04299998	.038
10. -STDEV-	.1095445	.4472135	.2073644	.1949359
11. -CI95-	.1361916	.5559999	.2578065	.2423548

Figure 4-5

ANALYSIS OF VARIANCE ON DIETS
WEIGHT GAIN (GMS) (NOTES)

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F STATISTIC	SIGNIFICANCE LEVEL
1. COLUMN	8.153499	3	2.717833	41.866	.0001
2. ROW	.3929987	4	.09824967		
3. RESIDUAL	.7790003	12	.06491669		
4. TOTAL	9.325498	19			

```
ROWS EXCLUDED: 6, 7, 8, 9, 10, 11

LEVENE'S STATISTIC = 2.273.  VARIANCES ARE HOMOGENEOUS AT P>0.05.

COLUMN 'DIET 1' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'DIET 2' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'DIET 3' IS NORMAL AT P>0.05 BY W-STATISTIC.
COLUMN 'DIET 4' IS NORMAL AT P>0.05 BY W-STATISTIC.


NULL HYPOTHESIS  H0: THE POPULATION MEANS ARE EQUAL.

INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.
```

Figure 4-6

TABLE 4-2. TWO-WAY ANALYSIS OF VARIANCE.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Value
Groups	$n\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2$	$k-1$	(Group SS)/(k − 1)	(Group MS)/(Residual MS)
Blocks	$k\sum_{j=1}^{n}(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})^2$	$n-1$	(Block SS)/(n − 1)	
Residual	$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})^2$	$(k-1)(n-1)$	(Residual SS)/ (k − 1)(n − 1)	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{\cdot\cdot})^2$	$nk-1$		

When there are missing values in a two-way design they are replaced with values which minimize the error sum of squares. Up to 10% of the data can be missing. If X_{ij} is empty it will be replaced with

$$X_{ij} = \frac{k\sum\bar{X}_{i\cdot} + n\sum\bar{X}_{\cdot j} - nk\bar{X}_{\cdot\cdot}}{(k-1)(n-1)}$$

When there is more than one missing value the above formula is applied iteratively until the formula holds for each case. The degrees of freedom for the residual and total sources of variation are reduced by the number of missing values.

4.7 Multiple Comparisons for Parametric Tests

When the initial results from the analysis of variance indicate that the true population means are not equal (i.e., the null hypothesis is rejected), it is usually of interest to determine specifically which pairs of samples are significantly different. This is accomplished by performing a multiple comparison test. The Newman-Keuls test is used when all possible pairwise comparisons are to be made. When one specific sample (called the "control") is to be compared to all other samples, Dunnett's test is employed. In the COMPARE dialogue you are asked if there is a control sample.

Both the Newman-Keuls and Dunnett procedures are known as multiple range tests. The different sample means are arranged in order of magnitude, and the various pairs of means are compared according to a prescribed sequence. Specifically, the largest mean is compared to the smallest first, then to the next smallest, etc. Then the next largest is compared with the smallest, next smallest, etc. This is continued until the two smallest means are compared.

The number of means between the two being compared plus one is known as the range of the comparison. An important principle of multiple range testing is that when a nonsignificant comparison is found, no further comparisons within that range are considered to be significant. For example, assume five ordered means, X1, X2, X3, X4 and X5. If X3 and X5 are found not to be signficantly different, then even without specifically performing the comparison (see the formula below), the comparisons X5 vs. X4 and X4 vs. X3 are considered not to be significantly different.

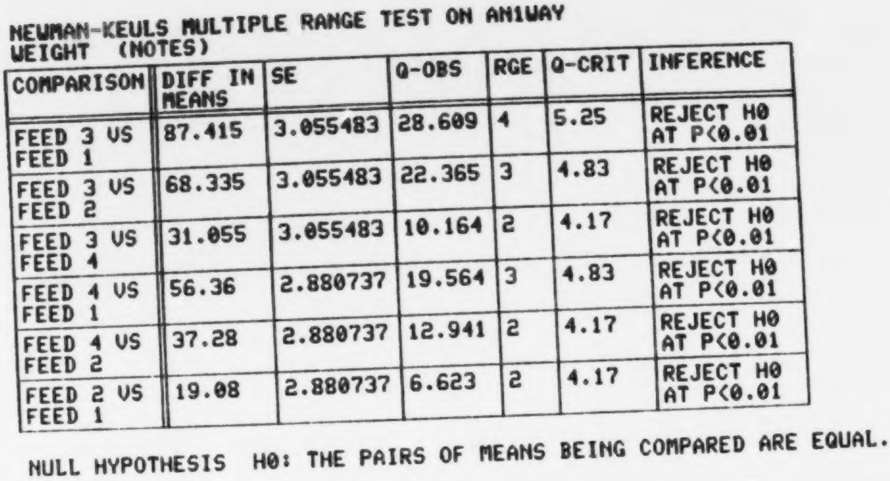
The calculated statistic (q) for the Student-Newman-Keuls test is

$$q = \frac{\overline{X}_A - \overline{X}_B}{SE}$$

where

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Tables of critical values are used which simultaneously take into account the total number of data in the samples and the range. The results of the Newman-Keuls multiple comparison for the data of Figure 4-3 are shown in Figure 4-7.



NEWMAN-KEULS MULTIPLE RANGE TEST ON AN1WAY
WEIGHT (NOTES)

COMPARISON	DIFF IN MEANS	SE	Q-OBS	RGE	Q-CRIT	INFERENCE
FEED 3 VS FEED 1	87.415	3.055483	28.609	4	5.25	REJECT H0 AT P<0.01
FEED 3 VS FEED 2	68.335	3.055483	22.365	3	4.83	REJECT H0 AT P<0.01
FEED 3 VS FEED 4	31.055	3.055483	10.164	2	4.17	REJECT H0 AT P<0.01
FEED 4 VS FEED 1	56.36	2.880737	19.564	3	4.83	REJECT H0 AT P<0.01
FEED 4 VS FEED 2	37.28	2.880737	12.941	2	4.17	REJECT H0 AT P<0.01
FEED 2 VS FEED 1	19.08	2.880737	6.623	2	4.17	REJECT H0 AT P<0.01

NULL HYPOTHESIS H0: THE PAIRS OF MEANS BEING COMPARED ARE EQUAL.

Figure 4-7

In this table COMPARE reports: what samples are being compared (e.g., the column heading or rowname for each sample); the difference in their means; the SE for the comparison; the calculated q statistic; the range of means for the test; the critical q-value; and the resulting inference. The significance level for the calculated q-value is the p-value used in the inference column. Critical values are available for the 0.05 and 0.01 significance levels. When a comparison is significant at the 0.01 level, the 0.01 critical value is listed in the results table. Otherwise the 0.05 critical value is listed.

For Dunnett's test the calculated q-value is

$$q' = \frac{\overline{X}_A - \overline{X}_{control}}{SE}$$

where

$$SE = \sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_{control}}\right)}$$

Again, s^2 is the residual mean square in the analysis of variance, and tables of critical values at the 0.05 and 0.01 level are used, where the total number of comparisons versus control are taken into account. An example of the results table for Dunnett's test is shown in Figure 4-8.

DUNNETT'S TEST ON AN1WAY
WEIGHT (NOTES)

COMPARISON	DIFF IN MEANS	SE	Q-OBS	RGE	Q-CRIT	INFERENCE
CONTROL VS FEED 2	-19.08	4.073977	4.683	4	3.43	REJECT H0 AT P<0.01
CONTROL VS FEED 4	-56.36	4.073977	13.834	4	3.43	REJECT H0 AT P<0.01
CONTROL VS FEED 3	-87.415	4.321106	20.23	4	3.43	REJECT H0 AT P<0.01

NULL HYPOTHESIS H0: THE PAIRS OF MEANS BEING COMPARED ARE EQUAL.

Figure 4-8

Before the results are displayed, you are also asked if you wish to see the full results table (as in Figure 4-7 and 4-8). If you respond "NO", only the results shown in Figure 4-9 will be displayed. You are also asked if you wish to see the results for nonsignifiant comparisons. If you respond "NO", only the data for significantly different means will be displayed. Remember when you display all comparisons that the results of comparisons made within a nonsignificant range should be viewed with caution.

NEWMAN-KEULS MULTIPLE RANGE TEST ON DIETS
WEIGHT GAIN (GMS) (NOTES)

COMPARISON	DIFF IN MEANS	INFERENCE
DIET 2 VS DIET 4	1.56	REJECT H0 AT P<0.01
DIET 2 VS DIET 1	1.42	REJECT H0 AT P<0.01
DIET 2 VS DIET 3	.56	REJECT H0 AT P<0.01
DIET 3 VS DIET 4	1.	REJECT H0 AT P<0.01
DIET 3 VS DIET 1	.86	REJECT H0 AT P<0.01
DIET 1 VS DIET 4	.14	ACCEPT H0 AT P>0.05

NULL HYPOTHESIS H0: THE PAIRS OF MEANS BEING COMPARED ARE EQUAL.

Figure 4-9

4.8 Kruskal-Wallis Test

The Kruskal-Wallis test extends the nonparametric two-sample tests to situations where there are three or more samples. This test is useful when the data demonstrate significant departures from normality. Recall, however, that the model is the same as introduced in Section 4.3 and the three assumptions described in Section 4.4 are in effect. Gross deviations from homogeneous dispersions will thus result in loss of accuracy. As in the case of the Wilcoxon-Mann-Whitney test (Section 3), all the data are assigned an overall rank, and the test is based on an evaluation of the rank sums for the different groups. To invoke the Kruskal-Wallis analysis, you again use the COMPARE command, specifying that you want to use the nonparametic test for data that are not blocked.

The quantities involved in the test for k samples are as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$C = 1 - \frac{\Sigma(t_i^3 - t_i)}{N^3 - N}$$

$$H_C = H/C$$

where Ri is the sum of ranks for the ith group, C is a correction for tied ranks, and t_j is the number of tied ranks of any group of ties. A table of critical values is used when k=3 and the sample sizes are less than or equal to 5. Otherwise, H_C is compared to the χ^2 distribution with k-1 degrees of freedom.

An example of the Kruskal-Wallis test for the data in Figure 4-10 is shown in Figure 4-11.

PONDS: PH LEVEL 8R X 8C (NOTES)

	1 POND 1	2 RANK(1)	3 POND 2	4 RANK(2)	5 POND 3	6 RANK(3)	7 POND 4	8 RANK(4)
1	7.68	1	7.69	2	7.74	13.5*	7.71	6*
2	7.7	3.5*	7.7	3.5*	7.75	16	7.71	6*
3	7.72	8	7.71	6*	7.77	18	7.74	13.5*
4	7.73	10*	7.73	10*	7.78	20*	7.79	22
5	7.73	10*	7.74	13.5*	7.8	23.5*	7.84	28
6	7.76	17	7.74	13.5*	7.81	26*	7.85	29
7	7.78	20*	7.78	20*	7.81	26*	7.87	30
8	7.8	23.5*	7.81	26*			7.91	31

STARRED RANKS INDICATE TIES

Figure 4-10

KRUSKAL-WALLIS TEST ON PONDS
PH LEVEL (NOTES)

	N	MEDIAN	RANK SUM
1. POND 1	8	7.73	93.
2. POND 2	8	7.735	94.5
3. POND 3	7	7.78	143.
4. POND 4	8	7.815	165.5

COLUMNS EXCLUDED: 2, 4, 6, 8

LEVENE'S STATISTIC = 4.68. VARIANCES ARE NOT HOMOGENEOUS AT P<.01.

NULL HYPOTHESIS H0: THE POPULATIONS ARE EQUAL.

KRUSKAL-WALLIS STATISTIC = 7.378 P = .061
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.

Figure 4-11

4.9 Friedman's Test

Friedman's test is used in situations where the two-way analysis of variance procedure is inappropriate on the basis of gross nonnormality or heterogeneity of variance. As in the two-way analysis of variance, the model for this test is

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

Again the three assumptions described in Section 4.4 apply, so that grossly heterogeneous dispersions are not appropriate. The data within each of the blocks is assigned a rank, and then rank sums for the samples are calculated. The test statistic A_c is

$$A_c = \frac{12 \sum_{i=1}^{k} R_i^2 - 3n^2k(k+1)^2}{nk(k+1) - \sum(t_j^3 - t_j)/(k-1)}$$

For certain combinations of sample size (n) and number of samples (k) the calculated statistic is compared to a table of critical values. These are: $k=3$, $n \leq 13$; $k=4$, $n \leq 8$; $k=5$, $n \leq 5$. Otherwise, this statistic is compared to the χ^2 distribution with k-1 degrees of freedom. When A_c meets or exceeds the critical value, the null hypothesis (i.e., that there are no differences among the populations from which the samples were drawn) is rejected. Figure 4-12 shows the analysis by Friedman's test on the data in Figure 4-5.

Note that there must be no missing data in the Friedman test. As a result the entire block in which an empty cell is found will be excluded by COMPARE. This will be reported along with the results table.

```
FRIEDMAN'S TEST ON DIETS
WEIGHT GAIN (GMS)  (NOTES)
+-----------++--+--------+-----------+
|           ||N |MEDIAN  |RANK SUM   |
+-----------++--+--------+-----------+
|1. DIET 4  ||5 |1.3     |7.         |
|2. DIET 1  ||5 |1.4     |8.         |
|3. DIET 3  ||5 |2.2     |16.        |
|4. DIET 2  ||5 |2.9     |19.        |
+-----------++--+--------+-----------+


LEVENE'S STATISTIC = 2.273.  VARIANCES ARE HOMOGENEOUS AT P>0.05.

NULL HYPOTHESIS  H0: THE POPULATIONS ARE EQUAL.

FRIEDMAN'S STATISTIC = 12.6   P = .006
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.
```

Figure 4-12

4.10 Nonparametric Multiple Comparison Tests

These procedures parallel the methodology of multiple comparisons discussed in Section 4.7 for the parametric cases; however, rank sums are used in place of means, and the standard errors are calculated in a different way. Table 4-3 shows the test statistics used for multiple comparison procedures based on the Kruskal-Wallis and Friedman tests. For certain combinations of sample size and number of samples the test statistic in column 1 is compared to a table of critical values. Each test has its own tables for the all-pairwise comparisons and for comparisons of all the treatments vs. a control. The second and third columns of Table 4-3 show the maximum sample sizes and number of samples for which tables of exact critical values are available on PROPHET.

TABLE 4-3. NONPARAMETRIC MULTIPLE COMPARISON STATISTICS.

Test	Small Sample Test Statistic	All Pairwise n, k	Control n, k	Large Sample Test Statistic	SE^2 All Pairwise*
Kruskal-Wallis	$\lvert R_A - R_B \rvert$	$n \leq 5, k = 3$	$n \leq 6, k = 3$	$\frac{\lvert R_A - R_B \rvert}{SE}$	$\frac{N(N+1)}{24}(\frac{1}{n_A} + \frac{1}{n_B})$
Friedman	$\lvert R_A - R_B \rvert$	$n \leq 15, k \leq 15$	$n \leq 18, k = 3$ $n \leq 5, k = 4$	$\frac{\lvert R_A - R_B \rvert}{SE}$	$\frac{N(k+1)}{12}$

*SE^2 for Control = 2[SE^2 for All Pairwise]

For larger sample sizes the test statistics are standardized as shown in column 4 of Table 4-3. The standard error used is shown in column 5 for the all-pairwise comparison case. When the comparisons are made with a control sample the standard error is increased by a factor of $\sqrt{2}$. The appropriate standardized test statistic is compared to a table of critical values for either the Newman-Keuls test or Dunnett's test using the range of the pair of samples being compared. The critical values will be approximate if the sample sizes are different in the Kruskal-Wallis test.

Examples are shown for multiple comparison using the Kruskal-Wallis test (Figure 4-13) and Friedman's test (Figure 4-14). As in performing multiple comparisons following the analysis of variance, you can display an abbreviated results table or only significant comparisons.

KRUSKAL-WALLIS TEST ON PONDS
PH LEVEL (NOTES)

COMPARISON	DIFF IN RANK SUMS	STD DIFF	PROB LEV	INFERENCE
POND 4 VS POND 1	9.0625	1.993484	.2772	ACCEPT H0 AT P>0.05
POND 4 VS POND 2	9.226562	2.029573	.2544	ACCEPT H0 AT P>0.05
POND 4 VS POND 3	20.10533	4.272621	.0006	REJECT H0 AT P<0.01
POND 3 VS POND 1	8.803571	1.870863	.3684	ACCEPT H0 AT P>0.05
POND 3 VS POND 2	8.894133	1.890109	.3528	ACCEPT H0 AT P>0.05
POND 2 VS POND 1	.1875	.0412445	.967	ACCEPT H0 AT P>0.05

LARGE SAMPLE APPROXIMATION
NULL HYPOTHESIS H0: THE POPULATIONS BEING COMPARED ARE EQUAL.

Figure 4-13

FRIEDMAN'S TEST ON DIETS
WEIGHT GAIN (GMS) (NOTES)

COMPARISON	DIFF IN RANK SUMS	CRIT DIFF	INFERENCE
DIET 2 VS DIET 4	12.	11.	REJECT H0 AT P<0.05
DIET 2 VS DIET 1	11.	11.	REJECT H0 AT P<0.05

EXACT METHOD
NULL HYPOTHESIS H0: THE POPULATIONS BEING COMPARED ARE EQUAL.

Figure 4-14

4.11 PROPHET Usage: TEST VARIANCES

The COMPARE command will automatically test whether the sample variances are homogeneous using Levene's test. But if you want to perform the test separately to see whether the data need be transformed before calling COMPARE, you can use the TEST VARIANCES command. The TEST VARIANCES command will also supply a summary of the variance and size for each sample.

TEST VARIANCES gives you the option of performing either Levene's test or Bartlett's test. Levene's test is very robust and can be used with data which are not normally distributed in order to test for homogeneity of dispersions. Bartlett's test is very powerful in tests for heteroscedasticity in samples whose populations are known to be normal; however, it behaves very poorly when the data are nonnormal.

TEST VARIANCES asks a series of questions (similar to COMPARE) to set up the test. You are first asked for the name of the data table, whether the data are organized by rows or columns, and if the analysis is to be performed on the entire table. This dialogue is illustrated at the top of Figure 4-15.

Levene's test employs the following test statistics (k = the number of samples):

$$Z_{ij} = | X_{ij} - \overline{X}_{i.} |$$

$$W_0 = \frac{\Sigma\, n_i (\overline{Z}_{i.} - \overline{Z}_{..})^2/(k-1)}{\Sigma\Sigma(Z_{ij} - \overline{Z}_{i.})^2/(N-k)}$$

The distribution of W_0 closely follows the F distribution with k-1 degrees of freedom. When W_0 exceeds the critical F, the null hypothesis, i.e., that the variances are homogeneous, is rejected.

```
>TEST VARIANCES<

ARE YOU COMPARING 2 OR MORE THAN 2 GROUPS (2 OR M)?M<
NAME OF TABLE?AN1WAY<
DATA FROM ROWS OR COLUMNS (R OR C)?C<
DO YOU WISH TO PERFORM THE ANALYSIS ON THE ENTIRE TABLE ?Y<

DO YOU WISH TO PERFORM LEVENE'S TEST (L) OR BARTLETT'S TEST (B) ?B

BARTLETT'S TEST FOR HOMOGENEITY OF VARIANCE ON AN1WAY: WEIGHT  (NOTES)
+-----------++----------+---+
|           || VARIANCE |DF |
+-----------++----------+---+
|1. FEED 1  || 46.34814 |4. |
|2. FEED 2  || 40.96289 |4. |
|3. FEED 3  || 37.28125 |3. |
|4. FEED 4  || 40.32715 |4. |
+-----------++----------+---+

NULL HYPOTHESIS H0: THE POPULATION VARIANCES ARE EQUAL.

BARTLETT'S STATISTIC = .039   P = .997
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.
```

Figure 4-15

Bartlett's test uses the following calculations:

$$B = (\ln s_p^2)\left(\sum_{i=1}^{k} \nu_i\right) - \sum_{i=1}^{k} \nu_i \ln s_i^2$$

$$\nu_i = n_i - 1$$

$$s_p^2 = \sum_{i=1}^{k} SS_i / \sum_{i=1}^{k} \nu_i$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k} \nu_i}\right)$$

$$B_c = B/C$$

The value of B_c is compared to the chi-squared distribution with k-1 DF.

When the variances are not equal it is often useful to try some form of data transformation in order to stabilize the variances. Transforming the data can have drawbacks, however; for example, the assumption that effects are additive may be violated by the transformation.

5. Frequency Anal.

5. Frequency Anal.

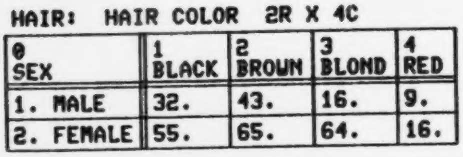
Section 5

FREQUENCY ANALYSIS

5.1 Contingency Tables

Often it is not possible to put a numerical value on an observation, but it is possible to put the data into categories. Data categorized "rats", "dogs", "redheaded", "lively", etc., are called nominal data. If you are able to put an ordering or ranking on the categories--e.g., "poor", "medium", and "good" -- then the data are called ordinal. Of course, ordinal data are also nominal data when the ordering is ignored.

Consider, for example, nominal data with two variables of interest: hair color (black, brown, blond, red) and sex (male, female). By counting the number of occurrences of any combination of categories of the two variables you can create a table such as that shown in Figure 5-1 (Zar, p. 59). Figure 5-1 shows, for example, that there are 32 males with black hair. Tables of this type are known as frequency tables or contingency tables. The table in this example is called a 2 x 4 contingency table because the variables create two rows by four columns.



HAIR: HAIR COLOR 2R X 4C

0 SEX	1 BLACK	2 BROWN	3 BLOND	4 RED
1. MALE	32.	43.	16.	9.
2. FEMALE	55.	65.	64.	16.

Figure 5-1

A primary question with nominal data is whether the two variables are independent of one another: that is, whether the knowledge of one variable gives any information about the other variable. An observation will be the categories for each variable on a particular subject, e.g., one observation is a male with red hair. The independence of data of this type can be analyzed by the X^2 goodness of fit test.

5.2 PROPHET Usage: CHISQ -- χ^2 Test for Independence

The public procedure CHISQ computes the χ^2 test of independence, the coefficient of contingency, and Cramer's V. To use CHISQ, you supply a two-way contingency table of cell frequencies for each row and column category. To invoke CHISQ, type the following:

```
RAC PUBLIC CHISQ
CALL CHISQ(contingency-table)
```

where contingency-table is a table of frequency data.

The program will ask if you wish to compute χ^2 for independence. You should respond with "YES" to test the whole table for independence. Otherwise, CHISQ performs a test for goodness of fit of a column of observed frequencies with a column of expected frequencies (see Section 5.6). Figure 5-2 shows the results of CHISQ for the data in Figure 5-1.

```
>RAC PUBLIC CHISQ<
>CALL CHISQ(HAIR)<

DO YOU WISH TO TEST FOR INDEPENDENCE?Y<
WOULD YOU LIKE TO SEE EXPECTED VALUES FOR EACH CELL?Y<

CHI SQUARE = 8.987   DEGREES OF FREEDOM = 3   SIG LEV = .029
COEFF OF CONTINGENCY = .1705         CRAMER'S V = .1731

EXPECTED VALUES FOR HAIR  2R X 4C
+-----------++------+------+-------+------+
|0          ||1     |2     |3      |4     |
|SEX        ||BLACK |BROWN |BLOND  |RED   |
+-----------++------+------+-------+------+
|1.         ||29.   |36.   |26.667 |8.333 |
+-----------++------+------+-------+------+
|2.         ||58.   |72.   |53.333 |16.667|
+-----------++------+------+-------+------+

WOULD YOU LIKE TO SEE CHI SQUARE PARTITIONED BY ROWS AND COLUMNS?Y<

PARTITIONED CHI SQUARE FOR HAIR  3R X 5C
+-------------++------+------+------+----+-----------+
|0            ||1     |2     |3     |4   |5          |
|SEX          ||BLACK |BROWN |BLOND |RED |CHI SQUARE |
+-------------++------+------+------+----+-----------+
|1. MALE      ||32.   |43.   |16.   |9.  |5.991      |
+-------------++------+------+------+----+-----------+
|2. FEMALE    ||55.   |65.   |64.   |16. |2.996      |
+-------------++------+------+------+----+-----------+
|3. CHI SQUARE||.466  |2.042 |6.4   |.08 |8.987      |
+-------------++------+------+------+----+-----------+
```

Figure 5-2

Expected Values

If you would like to see the expected values for each cell in the table, you respond "YES" to the question "WOULD YOU LIKE TO SEE EXPECTED VALUES?". After X^2 has been given, you will also be asked if you would like to see X^2 partitioned by rows and columns of the table. Respond "YES" to see a table with row-wise and column-wise contributions to X^2. These are the contributions each row and column have in the calculation of X^2. You will be warned if any of the expected frequencies are either less than 1.0 or more than 20% of the expected freuqencies contain values less than 5.0.

5.3 CHISQ Statistics

A. Interpreting chi-squared test of indepenence

Of the 300 people studied in our example (Figure 5-1), 25 were redheads. If we were to choose one person at random from the total group, the probability that the person is a redhead is 1/12. Similarly, the probability of that person being male would be 100/300, or 1/3. Under the null hypothesis that the two variables, hair color and sex, are independent, we would expect the probability of selecting a male redhead to be a product of the two probabilities, i.e., 1/36. Thus the expected number of male redheads is 300/36 = 8.33. That is, if the data are independent the proportion of males with red hair should be the same as the overall proportion of men; the proportion of redheads that are male should be the same as the overall proportion of redheads.

To generalize, the expected cell frequencies, F_{ij} depend only on the marginal totals (row and column total):

$$F_{ij} = \frac{R_i C_j}{N}$$

where R_i is the total in row i, C_j is the total in column j, and N is the overall total.

If the data are not independent, the number of male redheads would be either much smaller or much larger than expected.

To test whether the differences are greater than expected under the null hypothesis, we compare the observed frequency f_{ij} with the expected frequency F_{ij}, using the following formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

When the data are independent, this statistic follows the chi-squared distribution with (r-1)(c-1) degrees of freedom, where r is the number of rows and c the number of columns. Once we know (r-1)(c-1) of the cell frequencies and the marginal totals, the remaining c+r-1 frequencies can be ascertained.

Determining independence:

The greater the discrepencies are between the expected and the observed frequencies, the larger χ^2 becomes and the smaller is the probability that the data are independent.

The significance level is the probability of obtaining as large a χ^2 as observed under the null hypothesis. If this probability is small (<.05, say), we reject the null hypothesis and conclude that the two variables are related. The χ^2 test tells only whether the variables are independent or not. It does not indicate how strongly they are related.

Restrictions on chi-squared tests:

You should also note that if the value of any of the expected cell frequencies is less than 1, or if more than 20% of the expected cell frequencies contain values less than 5, the χ^2 is not valid. CHISQ reports this condition when it occurs.

One way to solve this problem is to look at the expected frequencies and see whether it makes sense to collapse two rows or columns by combining two categories. You may be able to correct the problem by collapsing enough categories. For example, the categories "red hair" and "blond hair" could be combined to form "not brown-haired". If, however, it makes no sense to combine categories, such as dead and alive, you should repeat the experiment so that the marginal totals (row or column totals) are large enough to obtain expected cell frequencies greater than 5.

What sample size to use is difficult to determine beforehand in cases where you cannot control the marginal totals. For example, if you wanted to know whether the survival of animals is independent of whether or not a drug is administered, it would not be possible to determine in advance the number of animals who will survive.

B. Contingency coefficient

The contingency coefficient is a measure of association between two variables where one or both are nominal scale data. It indicates the strength of the relationship between variables. It is computed from the χ^2 value found in the test of independence:

$$C = \left(\frac{\chi^2}{\chi^2 + N}\right)^{1/2}$$

When no relationship exists, C equals 0.0. The upper limit of C, however, depends on the number of rows and columns of the table. For example, in a 2 x 2 contingency table, the maximum value of C is 0.707. This maximum value will approach 1.0 as the table gets larger. For this reason, values of C should only be compared when they are obtained from tables of the same dimensions (i.e., similar experimental designs).

C. Cramer's V

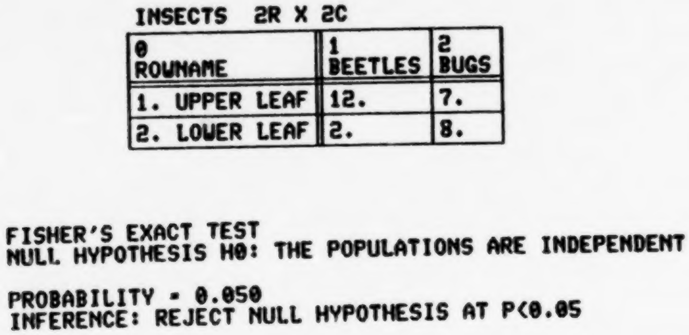
Another measure of association for nominal data based on the χ^2 statistic is Cramer's V. As with the contingency coefficient, it takes on the value 0.0 when the data are independent and increases in size as the strength of the relationship increases. But it can range from 0.0 to 1.0 for any size table by correcting by the minimum of the number of rows or the number of columns.

$$V = \left(\frac{\chi^2}{N \min(r-1, c-1)}\right)^{1/2}$$

These statistics have the same restrictions as the χ^2 in that the expected frequencies must be large enough. Even so, they are appropriate under a broad range of situations, where only nominal data are required.

5.4 PROPHET Usage: FISHER

When you test for independence in a 2 x 2 contingency table, X^2 is not accurate. Fisher's exact test should be used instead. For example, in Figure 5-3 the data consists of the number of a certain beetle species and a certain bug species, and whether they were found on the upper or lower surface of a leaf. (Zar, p. 292).



```
INSECTS  2R X 2C
0              1         2
ROWNAME        BEETLES   BUGS
1. UPPER LEAF  12.       7.
2. LOWER LEAF  2.        8.


FISHER'S EXACT TEST
NULL HYPOTHESIS H0: THE POPULATIONS ARE INDEPENDENT

PROBABILITY = 0.050
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05
```

Figure 5-3

The public procedure FISHER computes the exact significance level of an observed 2 x 2 contingency table under the null hypothesis that row and column classifications are equal. To use FISHER you type the following commands:

```
RAC PUBLIC FISHER
CALL FISHER(data-table)
```

FISHER will type the probability of observing such an extreme table as observed.

You can also assign the calculated probability value to a variable by the command:

```
SET variable-name TO FISHER(data-table)
```

The probability will also be typed out. The bottom of Figure 5-3 shows the results of Fisher's test.

Under the null hypothesis that the row and column classifications are independent, the probability of obtaining f_{11} from C_1 and f_{12} from C_2 are independent binomial probabilities. The probability of obtaining the observed table under the null hypothesis is

$$p = \frac{R_1!\, R_2!\, C_1!\, C_2!}{N!\, f_{11}!\, f_{12}!\, f_{21}!\, f_{22}!}$$

If the column and row totals are kept constant but the value of f_{11} is allowed to vary, several other tables are possible. Notice that with fixed marginal totals once f_{11} is picked all the remaining frequencies are determined. Now for each possible table the probability of obtaining that table can be computed. The observed table has a probability 0.02923 of occurring by chance when the data are independent. To see whether this is an unlikely event, it is compared to the probabilities of all the other tables. Tables with probabilities less than 0.02923 are considered more extreme tables than the observed table. If the total probability of obtaining a table at least as extreme as the observed table is less than, say, .05 you can reject the null hypothesis. But if the total probability of all extreme tables is large, then the observed table is not unusual under the null hypothesis and hence it is concluded that the data are independent.

5.5 PROPHET Usage: CROSSTAB and STATPAK

CROSSTAB

The public procedure CROSSTAB can be used to build a contingency table of up to 1000 cells from a data table of integer data. Each row of the data table corresponds to observations on several variables for a single subject. The different variables are in the columns. Each row of the contingency table will correspond to a different value in the first column of data specified and each column of the contingency table will correspond to a different value in the second column of data specified. Thus each cell will consist of the number of observations in the combined category. Optionally, you may control the range of values to be included in either variable. There is also an option to control for additional variables. You call CROSSTAB by typing the commands

```
RAC PUBLIC CROSSTAB
CALL CROSSTAB(data-table)
```

You will be asked to provide the number of the column containing the "row-variable" data and whether you want to specify a range of values to be included in the contingency table. If you choose to specify a range, you will be asked for them; otherwise the entire range of values in the column will be used.

Next, CROSSTAB asks for the number of the column containing the "column-variable" data. Here too you have an option of selecting a range of values.

You will also be asked if you want to control for any other variables in addition to the two primary variables. If you choose to do so, CROSSTAB asks for the range of values to be included in building the contingency table. That is, only those subjects which have data within the range of the additional variables will be included in the breakdown of the two primary variables. Once the contingency table is created, CROSSTAB will automatically call the public procedure STATPAK to perform the subsequent analysis.

Figure 5-4 shows the first few rows of a typical "raw-data-table". The cells in column 1, "SEX", are either 1 or 2 (male or female). The cells in column 2, "LESIONS" contain the number of lesions found in the animals; the cells in column 3 "AGE CAT" are numbers 1 to 5 and represent age categories.

LESIONS 100R X 3C

	1 SEX	2 LESIONS	3 AGE CAT.
1	1.	0.	1.
2	1.	1.	1.
3	1.	0.	2.
4	1.	0.	3.
5	1.	1.	2.
6	1.	1.	3.
7	1.	3.	5.
8	1.	0.	4.
9	1.	0.	5.
10	1.	1.	4.
11	1.	2.	5.
12	1.	1.	1.
13	1.	0.	1.
14	1.	0.	2.
15	1.	1.	3.
16	1.	1.	2.
17	1.	1.	3.
18	1.	2.	4.
19	2.	2.	1.
20	1.	2.	2.
21	1.	0.	4.
22	1.	2.	5.
23	1.	0.	1.
24	1.	1.	5.
25	1.	1.	1.
26	1.	0.	2.
27	1.	0.	3.
28	1.	1.	4.
29	1.	2.	5.
30	1.	0.	1.
31	1.	1.	2.
32	1.	1.	3.
33	1.	4.	4.
34	1.	2.	3.
35	1.	3.	5.
36	1.	1.	1.
37	1.	0.	2.
38	1.	1.	2.

Figure 5-4

Figure 5-5 shows a typical CROSSTAB dialogue to produce a contingency table as "SEX" vs. "LESIONS". By using an additional control, only those observations with AGE CAT equal to 4 and 5 are used.

```
>RAC PUBLIC CROSSTAB←
>CALL CROSSTAB(LESIONS)←

COLUMN # OF 1ST VARIABLE (ROW VAR.) = 2←

DO YOU WANT TO SPECIFY A RANGE OF VALUES FOR THE ROW VARIABLE?NO←

COL # OF 2ND VARIABLE (COL VAR.) = 1←

DO YOU WANT TO SPECIFY A RANGE OF VALUES FOR THE COLUMN VARIABLE?NO←

  DO YOU WANT TO CONTROL FOR OTHER VARIABLES ?YES←

    COLUMN # OF THE CONTROL = 3←
      MINIMUM VALUE OF CONTROL = 4←
      MAXIMUM VALUE OF CONTROL = 5←

  DO YOU WANT TO CONTROL FOR ANY OTHER VARIABLES ?NO←
```

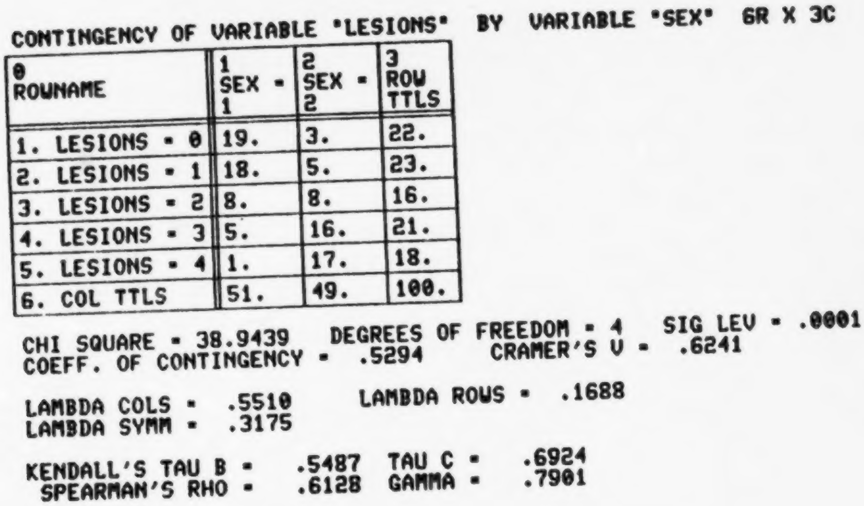
Figure 5-5

STATPAK

The public procedure STATPAK computes chi-squared, the coefficient of contingency, Cramer's V, lambda, Kendall's tau, Spearman's rho, and gamma on a contingency table. It will also collapse rows and columns of the table by adding rows or columns together if you wish and will optionally provide tables showing column, row and cell percentile breakdowns. To use STATPAK you type the commands

```
RAC PUBLIC STATPAK
CALL STATPAK(contingency-table)
```

where contingency-table is the name of an existing table. When you use the public procedure CROSSTAB to create a contingency-table, it will call STATPAK automatically.



```
CONTINGENCY OF VARIABLE "LESIONS"  BY  VARIABLE "SEX"  6R X 3C
+-----------------+-------+-------+------+
|0                |1      |2      |3     |
|ROWNAME          |SEX =  |SEX =  |ROW   |
|                 |1      |2      |TTLS  |
+-----------------+-------+-------+------+
|1. LESIONS = 0   |19.    |3.     |22.   |
|2. LESIONS = 1   |18.    |5.     |23.   |
|3. LESIONS = 2   |8.     |8.     |16.   |
|4. LESIONS = 3   |5.     |16.    |21.   |
|5. LESIONS = 4   |1.     |17.    |18.   |
|6. COL TTLS      |51.    |49.    |100.  |
+-----------------+-------+-------+------+

CHI SQUARE = 38.9439   DEGREES OF FREEDOM = 4   SIG LEV = .0001
COEFF. OF CONTINGENCY =  .5294     CRAMER'S V =  .6241

LAMBDA COLS =  .5510      LAMBDA ROWS =  .1688
LAMBDA SYMM =  .3175

KENDALL'S TAU B =   .5487  TAU C =   .6924
  SPEARMAN'S RHO =  .6128  GAMMA =   .7901
```

Figure 5-6

Options

First a table will be displayed showing column and row sums and the statistical measures described above will be displayed. The results of the data LESIONS in Figure 5-4 are shown in Figure 5-6. You will then be given three options. You can type:

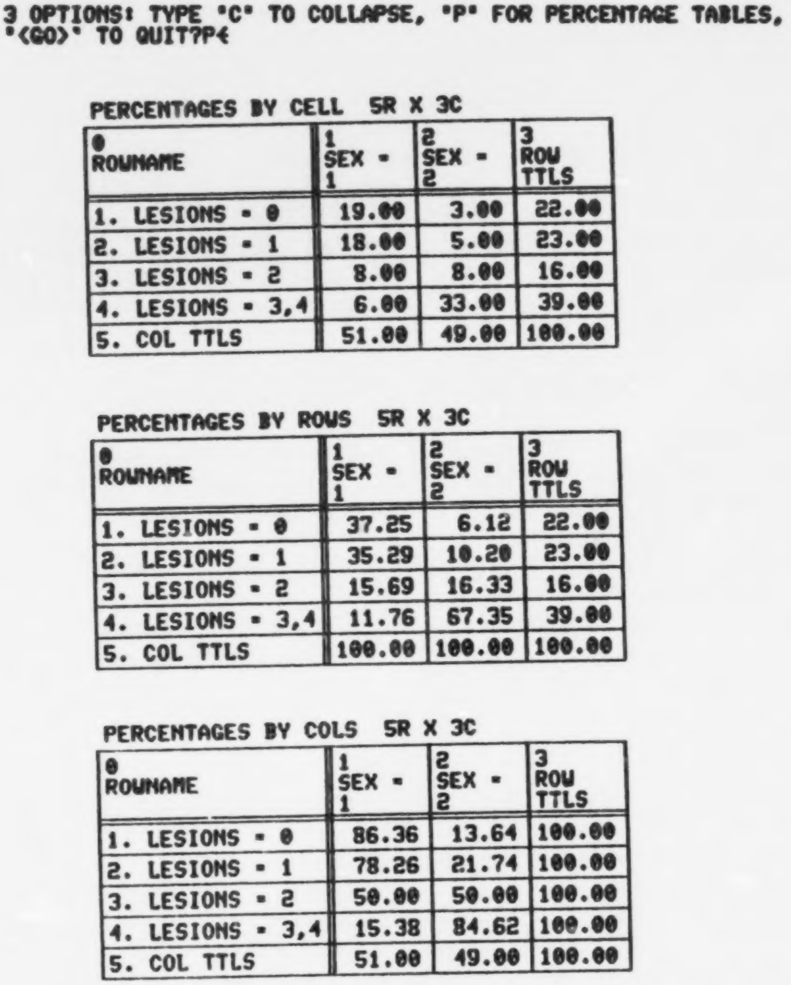
- "C" to collapse any two rows or columns
- "p" to have percentage tables created
- <GO> to exit from STATPAK

If you type "C", to collapse the table, you will be asked whether you want to collapse rows or columns. You answer with "R" or "C". The numbers of the two rows or columns are then asked for. Type them one at a time, following each by <GO>. The data in the two rows or columns will be summed, and a new display of the table and the statistics will be given. Figure 5-7 shows the results of collapsing rows 4 and 5.

```
3 OPTIONS: TYPE "C" TO COLLAPSE, "P" FOR PERCENTAGE TABLES,
"<GO>" TO QUIT?C€  COLLAPSE BY ROWS OR COLUMNS--TYPE "R" OR "C" ?R€
 COLLAPSE WHICH TWO ROWS?4€ ?5€


 CONTINGENCY OF VARIABLE "LESIONS"  BY  VARIABLE "SEX"  5R X 3C
+----------------+-----+-----+----+
|0               |1    |2    |3   |
|ROWNAME         |SEX =|SEX =|ROW |
|                |1    |2    |TTLS|
+----------------+-----+-----+----+
|1. LESIONS = 0  |19.  |3.   |22. |
|2. LESIONS = 1  |18.  |5.   |23. |
|3. LESIONS = 2  |8.   |8.   |16. |
|4. LESIONS = 3,4|6.   |33.  |39. |
|5. COL TTLS     |51.  |49.  |100.|
+----------------+-----+-----+----+

 CHI SQUARE = 37.6516   DEGREES OF FREEDOM = 3   SIG LEV = .0001
 COEFF. OF CONTINGENCY =  .5230     CRAMER'S V =  .6136

 LAMBDA COLS =  .5510     LAMBDA ROWS =  .2131
 LAMBDA SYMM =  .3636

 KENDALL'S TAU B =   .5537  TAU C =   .6648
  SPEARMAN'S RHO =   .6024  GAMMA =   .7952
```

Figure 5-7

If you answer "P" to the question, three tables will be displayed. The first shows the contribution of each cell, (expressed as a percent) to the total frequency of the table. The second table shows the contribution of each cell to its row total; the third,the contribution of each cell to its column total. Percentage tables are shown in Figure 5-8.



```
3 OPTIONS: TYPE "C" TO COLLAPSE, "P" FOR PERCENTAGE TABLES,
"<GO>" TO QUIT?P<

        PERCENTAGES BY CELL  5R X 3C
        +--------------------+--------+--------+--------+
        |0                   |1       |2       |3       |
        |ROWNAME             |SEX =   |SEX =   |ROW     |
        |                    |1       |2       |TTLS    |
        +--------------------+--------+--------+--------+
        |1. LESIONS = 0      |   19.00|    3.00|   22.00|
        |2. LESIONS = 1      |   18.00|    5.00|   23.00|
        |3. LESIONS = 2      |    8.00|    8.00|   16.00|
        |4. LESIONS = 3,4    |    6.00|   33.00|   39.00|
        |5. COL TTLS         |   51.00|   49.00|  100.00|
        +--------------------+--------+--------+--------+

        PERCENTAGES BY ROWS  5R X 3C
        +--------------------+--------+--------+--------+
        |0                   |1       |2       |3       |
        |ROWNAME             |SEX =   |SEX =   |ROW     |
        |                    |1       |2       |TTLS    |
        +--------------------+--------+--------+--------+
        |1. LESIONS = 0      |   37.25|    6.12|   22.00|
        |2. LESIONS = 1      |   35.29|   10.20|   23.00|
        |3. LESIONS = 2      |   15.69|   16.33|   16.00|
        |4. LESIONS = 3,4    |   11.76|   67.35|   39.00|
        |5. COL TTLS         |  100.00|  100.00|  100.00|
        +--------------------+--------+--------+--------+

        PERCENTAGES BY COLS  5R X 3C
        +--------------------+--------+--------+--------+
        |0                   |1       |2       |3       |
        |ROWNAME             |SEX =   |SEX =   |ROW     |
        |                    |1       |2       |TTLS    |
        +--------------------+--------+--------+--------+
        |1. LESIONS = 0      |   86.36|   13.64|  100.00|
        |2. LESIONS = 1      |   78.26|   21.74|  100.00|
        |3. LESIONS = 2      |   50.00|   50.00|  100.00|
        |4. LESIONS = 3,4    |   15.38|   84.62|  100.00|
        |5. COL TTLS         |   51.00|   49.00|  100.00|
        +--------------------+--------+--------+--------+
```

Figure 5-8

Even though STATPAK calculates many statistics, not all of them are appropriate at one time. Spearman's rho, Kendall's tau, and gamma can be correctly applied only to data which can have some ordering placed on them. The remaining statistics can be applied to either nominal or ordinal data, but are not very discriminating for ordinal data. Only one of the lambdas (row, column or symmetrical) makes sense for any one data set. Therefore, it is up to you to decide which statistics can legitimately be applied to your data.

5.6 STATPAK Statistics

A. Statistics for nominal data

You can test whether two nominal scale variables are independent of each other by the χ^2 test of independence described in Section 5.3. The coefficient of contingency and Cramer's V are measures of the amount of association between two nominal variables. These statistics are discussed in Section 5.3. In addition, lambda will measure the association between two variables when one is dependent on the other (as in regression).

B. Statistics for ordinal data

If there is an ordering on the data, we can measure the direction as well as the degree of the association between two variables. There are three major measures of association for ordinal data: Kendall's tau, gamma and Spearman's rho. These measures are similar to the Pearson correlation coefficient on interval data. When parametric statistics seem inappropriate, or your interest is in the correspondence between pairs of ranks instead of the correspondence between pairs of magnitudes, then interval data may be ranked and these three measures can be used.

C. Lambda

Lambda is a measure of association between two qualitative variables (nominal scale). The symmetric lambda, lambda columns and lambda rows measure the improvement in the ability to predict the value of the dependent variable given the knowledge of the independent variable: in other words, how much better you can predict which row an observation came from when you know the column than when you do not know the column. Lambda rows is a measure of the prediction of the rows given the columns. Lambda cols is the measure of association of the columns when the rows are the independent variable. Lambda symmetric is the measure of association between the two variables without assuming one of the variables is a dependent variable. It is a combination of the two other lambdas.

The lambdas are based on the assumption that the best strategy for prediction is to choose the most frequent (model) category. The idea is to compute the relative reduction in error by this prediction method, using the knowledge about the independent variable. For example, consider lambda columns, where the columns contain the dependent variable and you want to know whether you can predict which column the observations will be in. If the row and column variables are independent, then knowing in which row the data lie is no help. The best prediction is always to pick the column with the most observations. In Figure 5-9, column 1 is the column with the largest marginal total. The number of incorrect observations is the total number of observations minus the number of observations in the largest column marginal:

$$E_1 = N - \max_j C_j$$

But if the variables are related, then knowing which row the data come from will influence the choice of the column. The prediction rule is then to pick the column with the most observations in a particular row. Therefore, in row 1 column 2 would be chosen, since 27 is the largest number in row 1. Similarly columns 3, 1, and 2 would be chosen in rows 2, 3, and 4.

LAMBDA 5R X 5C

0 ROWNAME	1 C1	2 C2	3 C3	4 C4	5 TOTAL
1. R1	15.	27.	3.	8.	53.
2. R2	0.	5.	24.	10.	39.
3. R3	37.	9.	1.	6.	53.
4. R4	7.	13.	0.	4.	24.
5. TOTAL	59.	54.	28.	28.	169.

Figure 5-9

The total error in this prediction rule would be the sum of the incorrect choices in each row:

$$E_2 = \sum_{i=1}^{r} (R_i - \max_j f_{ij}) = N - \sum_{i=1}^{r} \max_j f_{ij}$$

In comparing the two prediction methods you are interested in the relative reduction in error. It is clear that since the row information is not being used in the first rule, the E_1 is larger than E_2 and so

$$\text{lambda cols} = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{r} \max_j f_{ij} - \max_j C_j}{N - \max_j C_j}$$

Thus fewer errors in predicting the columns will be made by using the knowledge of the rows than by using the column marginals only. Similarly if the dependent variable was the row then given the columns, rows 1, 2, 3 would be appropriate predictions for columns 1, 2, 3, 4. E_1 and E_2 would be adjusted to reflect this and thus

$$\text{lambda rows} = \frac{\sum_{j=1}^{c} \max_i f_{ij} - \max_i R_i}{N - \max_i R_i}$$

Notice that the prediction rules are somewhat different when columns are the independent variable and when rows are the independent variable; i.e., given col 4, row 2 would be the predicted outcome, whereas given row 2, you would predict col 3. Thus the difference in the two lambdas.

When there is no sensible way of designating either rows or columns as the dependent variable the lambda symmetric would be used. It is a composite measure of association reflecting the relative reduction in error in predicting both variables.

$$\text{lambda symmetric} = \frac{\sum_{i=1}^{r} \max_{j} f_{ij} + \sum_{j=1}^{c} \max_{i} f_{ij} + \max_{j} C_j - \max_{i} R_i}{2N - \max_{i} C_j - \max_{i} R_i}$$

Lambda may range from 0 to 1. A value of 0 indicates no reduction in error and hence no association between the variables. A value of 1 is a perfect prediction. It should be noted that lambda uses just one of several kinds of prediction rules. It is not numerically comparable to other measures of association.

D. Spearman's rho

Spearman's rho, r_s , is comparable to Pearson's correlation coefficient, r, but it is computed on the ranks of the data. If X_i are the ranks on the first variable and Y_i are ranks on the second variable, then

$$r = r_s = \frac{\sum X_i^2 + \sum Y_i^2 - \sum d_i^2}{2\sqrt{\sum X_i^2 \sum Y_i^2}}$$

If there are no differences in the ranks of the two variables, $r_s=1$ and the variables are said to be perfectly correlated. If the ranks of the two variables are exactly reverse then $r_s = 1$ (perfect negative correlation). When there is no correspondence between pairs of ranks r_s should be zero or near zero.

When two or more observations are the same for a variable the average rank is used in the computation of rho. A correction is introduced in computing ΣX_i and ΣY_i because the ties will inflate the value of rho.

Let t_x be the number of ties with a given rank on the variable X; then

$$\sum X_i^2 = \frac{N^3 - N - \sum (t_x^3 - t_x)}{12}$$

where the summation sign on the right is summing over each t_x corresponding to different ranks. A similar expression is written for the Σy_i variable.

Although the correlation between ranks of interval data will be higher in general than the correlation between the original interval data, this correlation will be less discriminating than that based on the observed measurements. By treating all differences on adjacent measurements as identical by ranking, a simplicity and regularity is imposed on the data which in fact it does not possess.

E. Kendall's tau

Whenever the number of observations exceeds the number of categories some ties are inevitable. Rho is inappropriate when there are large numbers of ties. Kendall's tau is a better measure of association. Suppose the data had been ranked or divided into order categories. Now consider a pair of observations, i and j on variables X and Y. If the relative ranking of the two observations is the same on the first variable X and the second variable Y, then the pair is concordant; otherwise they are discordant. For example, if $X_i<X_j$ and $Y_i<Y_j$ the observations are concordant. If either the ranks on the X's or on the Y's are the same the pair is ignored. Let P be the numbers of concordant pairs and let Q be the numbers of discordant pairs, when all possible pairs of observations are looked at. When all the pairs are concordant then P would equal N(N-1)/2. Kendall's tau is the difference in concordant or discordant pairs

$$\tau_b = \frac{P - Q}{N(N - 1)/2}$$

Since P-Q is exactly the number of interchanges between neighbors which is required to transform the ranking at one variable to the ranking of the other variables, you could think of tau as a "coefficient of disarray".

When there are tied variables, tau is corrected by the following formula:

$$\tau_b = \frac{P - Q}{\sqrt{\tfrac{1}{2}N(N - 1) - T_x}\ \sqrt{\tfrac{1}{2}N(N - 1) - T_y}}$$

The number of tied observations on X and Y are t_x and t_y and

$$T_x = \tfrac{1}{2}\sum t_x(t_x - 1)$$

$$T_y = \tfrac{1}{2}\sum t_y(t_y - 1)$$

τ_b is most appropriate with data that have the same number of categories on X as on Y. When this is not true then τ_c is more appropriate.

$$\tau_c = \frac{2m(P - Q)}{N^2(m - 1)}$$

where m is the minimum number or rows or columns.

F. Gamma

Gamma is similar to Kendall's tau, but it makes no adjustment for ties or the number of the data.

$$\gamma = \frac{P - Q}{P + Q}$$

where P and Q are defined as for Kendall's tau. Gamma is based on the relative reduction of error in predicting the order of a pair of observations on one variable given the ordering on the other variable. If there is no association between the two variables then simply guessing the order of a variable would work about half the time. Thus,

$$E_1 = \frac{P + Q}{2}$$

But if the knowledge of the order of variable X is used to guess the order on the Y variable then you would be wrong min(P,Q) times, i.e., E_2 equals min(P,Q). Therefore,

$$\gamma = \frac{E_1 - E_2}{E_1} = \frac{P - Q}{P + Q}$$

Thus the relative reduction in error by using the order on one variable is gamma.

Gamma and tau do not accurately measure association if the relationship changes abruptly. For example, in an extreme case, suppose there was perfect positive correlation for pairs in the lower range of values and perfect negative association for pairs in the upper range. Then gamma and tau would be exactly 0 since P=Q.

Gamma, tau, and rho are all measures of association which take on values from -1 to 1. If there is perfect positive correlation then they will equal 1. If there is perfect negative correlation then they will equal -1. If there is no correlation they will equal 0. But they all have different underlying scales and so numerically they are not directly comparable. But they do use the same amount of information in the data and thus have the same power to detect the existence of association.

5.7 PROPHET Usage: CHISQ -- χ^2 Test for Goodness of Fit

The χ^2 test can also be used to test whether the population from which the sample was drawn conforms to a certain theoretical distribution. For example, suppose you have data for peas classified as yellow or green and smooth or wrinkled and you want to test whether the data fall into a ratio of 9:3:3:1 for yellow-smooth, yellow-wrinkled, green-smooth, and green-wrinkled (Zar, p. 46). Figure 5-10 shows the data and the expected frequencies with the above ratios.

PEAS: FEQUENCIES OF TYPES OF PEAS 5R X 2C

0 ROWNAME	1 OBSERVED	2 EXPECTED
1. YELLOW-SMOOTH	152.	140.625
2. YELLOW-WRINKLED	39.	46.875
3. GREEN-SMOOTH	53.	46.875
4. GREEN-WRINKLED	6.	15.625
5. TOTAL	250.	250.

```
>RAC PUBLIC CHISQ<
>CALL CHISQ(PEAS)<

DO YOU WISH TO TEST FOR INDEPENDENCE?N<
    GIVE NUMBER OF COLUMN CONTAINING OBSERVED FREQUENCIES?1<
    GIVE NUMBER OF COLUMN CONTAINING COMPARISON FREQUENCIES?2<

CHI SQUARE = 8.972   DEGREES OF FREEDOM = 4   SIG LEV = .062
```

Figure 5-10

The public procedure CHISQ can be used to test whether it is reasonable to assume the data follow some proposed distribution. You supply a table where one column contains observed frequencies and a second column contains expected frequencies. To use CHISQ, you type the following commands

```
RAC PUBLIC CHISQ
CALL CHISQ (data-table)
```

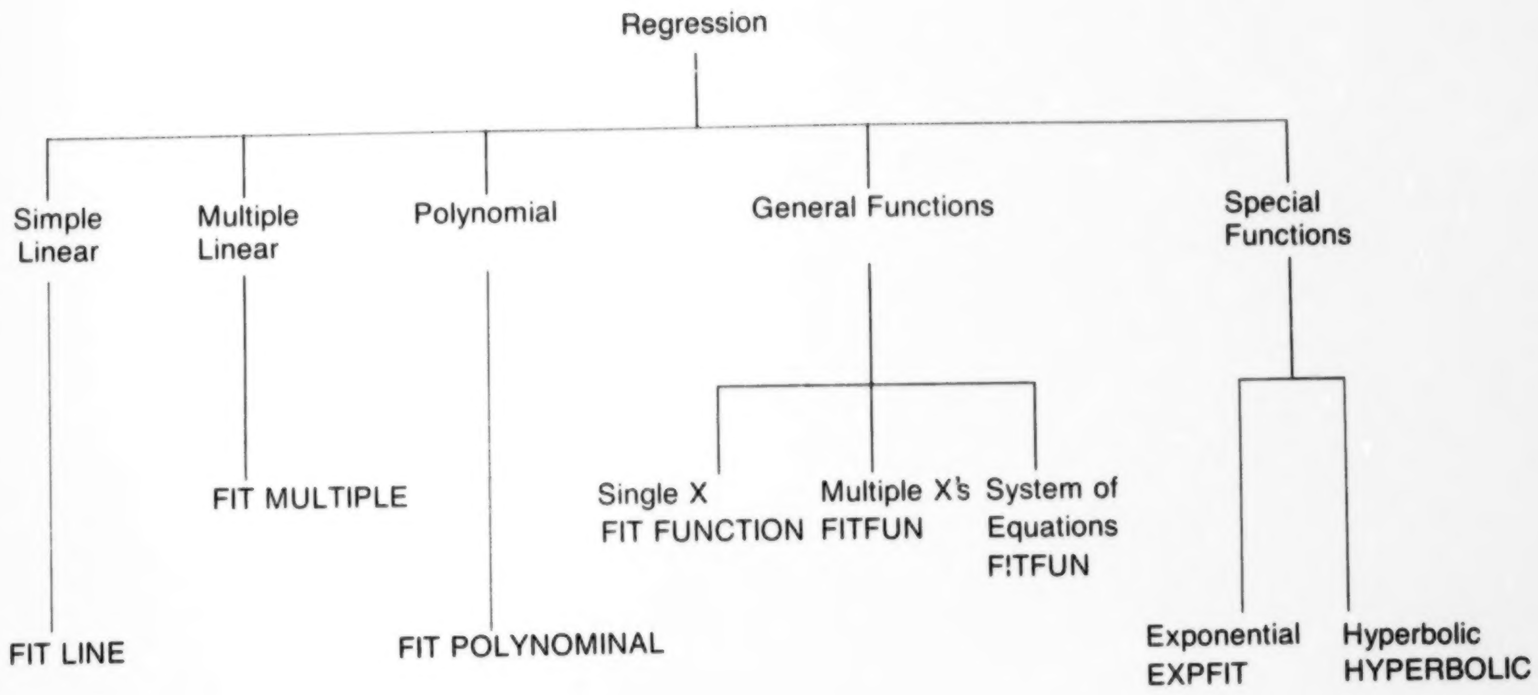
First you will be asked whether you want to test for independence (see Section 5.2). You should repond "NO".

You will be asked for the number of the column containing the observed data and for the number of the column containing the expected value data. The analysis is shown in Figure 5-8.

If it is possible to determine the expected frequency of categories, the χ^2 goodness of fit can be applied. If the χ^2 value is large enough to be significant, you reject the null hypothesis and assume the fit is not good.

The statistic is computed as

$$\chi^2 = \sum_{i=1}^{r} \frac{(f_i - F_i)^2}{F_i}$$


Regression
Simple Linear
Multiple Linear
Polynomial
General Functions
Special Functions
FIT LINE
FIT MULTIPLE
FIT POLYNOMINAL
Single X
FIT FUNCTION
Multiple X's
FITFUN
System of Equations
FITFUN
Exponential
EXPFIT
Hyperbolic
HYPERBOLIC

Section 6

REGRESSION ANALYSIS

6.1 Introduction

Previous sections of this manual have discussed statistical techniques for analysis of a single variable. Many experimental situations are, however, better described by considering simultaneously the values of two variables which are related by some mathematical function. For example, in a study on rats the magnitude of the response in blood presure is related to the dose of a vasoactive drug; in a treadmill experiment, the increase in heart rate is a function of the level of exercise. The goal of regression analysis is to determine the functional form, or model, that best expresses the values of the dependent variable (called Y), in terms of the independent variable (called X). In the second example above, heart rate is the dependent variable, since the response depends upon the level of exercise. The values of the independent variable are assumed to be fixed, with no error, and not to depend on the values of the Y-variable.

There are several common functional relationships which figure prominently in biostatistical applications. Among these are the simple linear, multiple linear, polynominal, and exponential models:

Simple Linear: $Y = \alpha + \beta X$

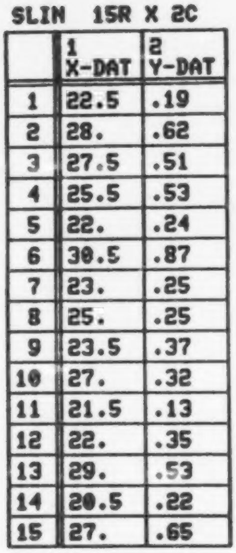
Multiple Linear: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

Polynomial: $Y = \beta_o + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$

Exponential: $Y = \alpha e^{\beta X}$

In each of the above equations, the mathematical relationship between X and Y is determined by the values of the parameters (α, β_0, $\beta_1 \ldots$). It is important to keep in mind that the regression analysis cannot determine the form of the mathematical function automatically. It is up to you to choose the form of the function in advance and then let the regression analysis determine the values of the parameters in the model. By assessing how well the model explains or predicts the values of Y for each value of X, by looking at the residuals (deviations in the observed values of Y from the values of Y predicted by the equation), and by checking the goodness of fit statistics, you can determine whether a different functional form should be tried.

Often the first step in regression analysis is to examine a plot of the data (called a scatter plot), with the independent variable (X) plotted on the abscissa, and the dependent variable (Y) plotted on the ordinate. A graph of the data of Figure 6-1 is plotted in Figure 6-2.



SLIN 15R X 2C

	1 X-DAT	2 Y-DAT
1	22.5	.19
2	28.	.62
3	27.5	.51
4	25.5	.53
5	22.	.24
6	30.5	.87
7	23.	.25
8	25.	.25
9	23.5	.37
10	27.	.32
11	21.5	.13
12	22.	.35
13	29.	.53
14	20.5	.22
15	27.	.65

Figure 6-1

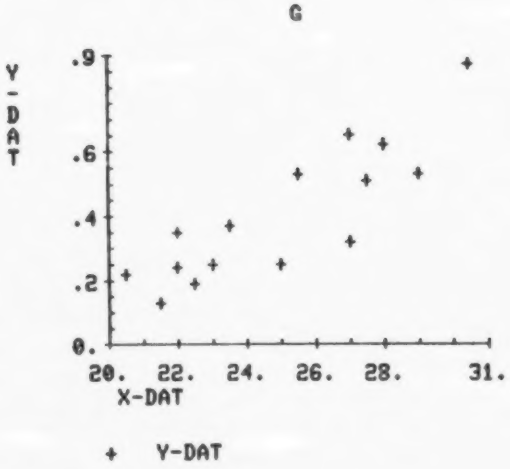


Figure 6-2

When an *a priori* model is not available, preliminary inspection of the data from the scatter plot is the best way of determining the form of the equation on which to base the initial analysis. It must be emphasized that (unless there are overriding theoretical considerations) any model must be considered tentative during the early stages of regression analysis. The model must be critically examined at several stages, and it should be changed if the results of the analysis do not support its validity.

6.2 Principles and Assumptions of the Regression Technique

In any regression analysis, determining an appropriate functional relationship does not imply that all observed values of Y are exactly related to the corresponding values of X by the resulting equation. But it is assumed that the true mean (or expected value) of Y corresponding to a given value of X is related by the prescribed function. While it is assumed that the independent variable X is measured without error (that is, it is under the control of the investigator), the dependent variable Y is a random variable. For each value of X_i the corresponding value of Y_i is assumed to be

$$Y_i = f(X_i) + \epsilon_i$$

where f(X) is some mathematical function and ε_i is the random error. The errors are assumed to be mutually independent (and therefore uncorrelated) and normally distributed with a mean equal to zero and unknown variance σ^2. Since X_i is assumed to be fixed, the distribution of Y_i must also be normal with mean $f(X_i)$ and variance equal to σ^2 for each value of X. That is, the variance is constant and independent of the magnitude of X. If you were to observe many values of Y at X_i, you expect the Y's to be distributed so that they cluster about $f(X_i)$ and trail off in frequency away from $f(X_i)$. Figure 6-3 shows the assumed distributions of Y at X_1, X_2 and X_3.

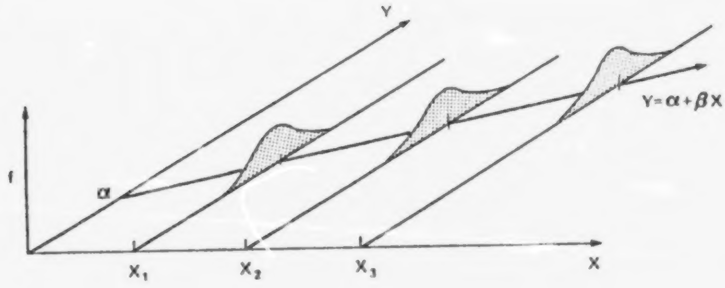


Figure 6.3

6.3 PROPHET Usage: Introduction

There are several procedures available in PROPHET that enable you to fit data to a variety of functional forms. These are shown on the roadmap facing page 6-1, and each will be discussed in more detail in the following sections.

In Section 6.4 the analysis of the simple linear model is discussed. Other subjects, such as the analysis of variance approach to regression, weighted least squares, and analysis of residuals, are also introduced in that section, but their general principles apply to other models as well.

The FIT LINE, FIT POLYNOMIAL, and FIT FUNCTION commands are similar in terms of the input you supply and the output that is produced. In all three data can be specified from columns of a table or a curve from a graph, you have the option of performing weighted regression analysis, and results are presented in a parameter table and an analysis of variance table. In addition, you can add the calculated regression curve to a graph and perform a variety of forms of graphical and tabular residual analysis.

FIT FUNCTION and FITFUN are PROPHET procedures available for fitting the dependent variable Y to general linear and nonlinear functions of the independent variable, including any combination of powers, logarithms, trigonometric functions, hyperbolic functions, etc. FIT FUNCTION is used for fuctions of a single X-variable. FITFUN is used in fitting data to functions with many X-variables and also can be used for fitting a system of functions simultaneously.

PROPHET also provides three Public Procedures for performing other forms of regression analysis: MULREG performs multiple linear regression; EXPFIT performs peeling for exponential curve fitting; and HYPERBOLIC fits data to a hyperbolic model.

6.4 Simple Linear Regression

A. General form

The simplest functional relationship between two variables is the linear first order equation:

$$Y = \alpha + \beta X$$

Graphically, this is simply a straight line. According to this model, for every unit increase in X, Y will increase by β units, and when X=0, Y= α.

As mentioned previously, the objective in regression analysis is to determine estimates for α and β (call them a and b), such that the resultant equation represents the "best fit" of the line to the observed data. The usual criterion for best fit employs the concept of "least squares estimation". For each value of X_i there is a corresponding observed value, Y_i, and a value predicted by the equation, $\hat{Y}_i$ (called Y hat). The method of least squares is designed to determine values of the coefficients of the regression equation, such that the sum of squares of the vertical deviations of $(Y_i-\hat{Y}_i)$ is a minimum.

The best estimate of β, the true regression coefficient, is b, which is given by:

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

The intercept is the value of Y when X=0. Since the pair of data points $\overline{X}$ and $\overline{Y}$ always lie on the regression line, the best estimator of α is a, estimated from the two sample means:

$$a = \overline{Y} - b\overline{X}$$

Notice that both a and b are necessary to specify the regression line uniquely.

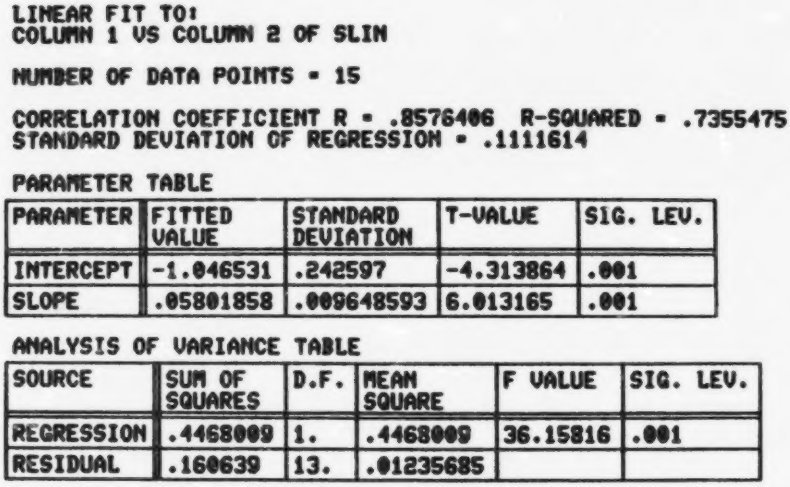
B. Analysis of Variance

An analysis of variance tests the precision of the estimated regression. The total variability of the data is described by three sum of squares (SS) terms as follows:

$$\Sigma (Y_i - \overline{Y})^2 = \Sigma (Y_i - \hat{Y}_i)^2 + \Sigma (\hat{Y}_i - \overline{Y})^2$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

The calculation of these quantities is shown in Table 6-1. Notice that the total sum of squares does not include any information about the X's. The corresponding mean square terms are calculated by dividing by the appropriate degrees of freedom (DF). The residual mean square, $s^2_{y.x}$ is the variance of Y after taking into account its dependence on X. Figure 6-4a shows the goodness of fit statistics of the data in Figure 6-1. Figure 6-4b shows the graph of the fit and Figure 6-4c shows the table of residuals.



```
LINEAR FIT TO:
COLUMN 1 VS COLUMN 2 OF SLIN

NUMBER OF DATA POINTS = 15

CORRELATION COEFFICIENT R = .8576406  R-SQUARED = .7355475
STANDARD DEVIATION OF REGRESSION = .1111614
```

PARAMETER TABLE

PARAMETER	FITTED VALUE	STANDARD DEVIATION	T-VALUE	SIG. LEV.
INTERCEPT	-1.046531	.242597	-4.313864	.001
SLOPE	.05801858	.009648593	6.013165	.001

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F VALUE	SIG. LEV.
REGRESSION	.4468009	1.	.4468009	36.15816	.001
RESIDUAL	.160639	13.	.01235685		

Figure 6-4a

```
WOULD YOU LIKE TO SEE A GRAPH OF THE FIT?
```

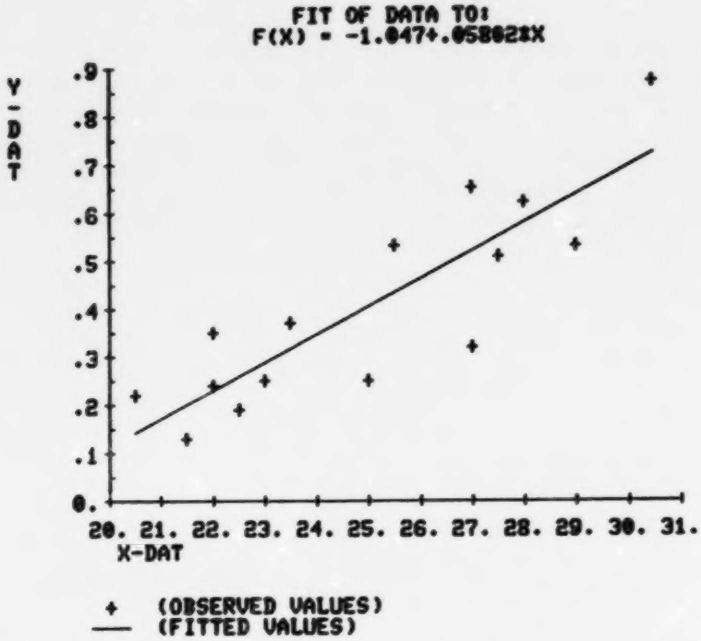


Figure 6-4b

```
WOULD YOU LIKE TO SEE A TABLE OF RESIDUALS?
```

TABLE OF RESIDUALS 15R X 4C

	1 X-DAT	2 Y-DAT (OBS)	3 (D) Y-DAT (FITTED)	4 (D) RESIDUAL
1	22.5	.19	.2588871	-.06888705
2	28.	.62	.5779892	.04201076
3	27.5	.51	.54898	-.03897996
4	25.5	.53	.4329428	.09705721
5	22.	.24	.2298778	.01012225
6	30.5	.87	.7230357	.1469643
7	23.	.25	.2878963	-.03789634
8	25.	.25	.4039335	-.1539335
9	23.5	.37	.3169056	.05309437
10	27.	.32	.5199707	-.1999707
11	21.5	.13	.2008685	-.07086847
12	22.	.35	.2298778	.1201222
13	29.	.53	.6360078	-.1060078
14	20.5	.22	.1428499	.07715011
15	27.	.65	.5199707	.1300293

Figure 6-4c

C. F-ratio

The ratio of the regression mean square to the residual mean square follows the F-distribution. This F-ratio measures how well the regression explains the variability of the data relative to the unexplained (residual) error. It is thus a test of the null hypothesis that the data are simply randomly distributed, with no systematic variation, versus the alternative hypothesis that a significant relationship between X and Y exists. The larger the F-value, the more significant the fit. Nonsignificant F-values may result from inadequacy of the model or excessive error in the observations.

D. Standard error of regression

The square root of $s^2{}_{y \cdot x}$, known as the "standard error of regression" or "standard deviation of regression", indicates the overall accuracy of the fitted line, i.e., how well the model predicts or explains the observed values of Y.

E. Correlation coefficient squared

The proportion of total variability that is explained by the regression is known as the correlation coefficient squared, R^2, defined as

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}}$$

Considering the two extreme cases of R^2 provides some insight into its meaning. When $R^2=0$, it implies that regression SS = 0. That is, none of the variability of the data may be attributed to the regression; it is entirely due to error. This is equivalent to saying that the two variables are entirely unrelated. Conversely, when $R^2=1$ or -1, all of the variability is due to the regression, there is no error or residual variability, and the X and Y variables are perfectly correlated.

F. Standard errors of slope and intercept

The calculated slope of the regression line, b, expresses the dependence of the observed Y-values on their associated X-values. What is really desired, however, is to determine the functional relationship in the population from which the observed samples were drawn. If there is no dependence of Y on X in the true population, then $\beta=0$. It is possible (in fact likely) that the sample drawn from the population will yield a value of b which is not equal to zero. Thus, it is important to test the null hypothesis that $\beta=0$ against the alternative hypothesis that $\beta \neq 0$. The standard error of the slope (b) and intercept (a) are calculated as follows:

$$SE(b) = \sqrt{\frac{s^2_{y \cdot x}}{\Sigma(X_i - \bar{X})^2}}$$

and

$$SE(a) = \sqrt{s^2_{y \cdot x}\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)}$$

In the usual fashion, the t-statistic, calculated as the quotient of the estimated parameter value and its standard error, may be used to test the null hypothesis that the population slope (or intercept) is equal to 0.

G. Analysis of residuals

The residuals from a regression analysis are defined as the set of differences $Y_i - \hat{Y}_i$; that is, they are the individual differences between what is observed, and what is predicted by the equation. It is a fundamental assumption in regression analysis that the residuals are independent (which implies that they are uncorrelated) and normally distributed with mean zero and constant variance for every value of X. The usual approach is to plot the value of each residual against its corresponding X, in order to look for obvious trends or abnormalities. If the basic assumptions hold, the residuals should be normally distributed about zero for all values of X. If there is a significant increase in the variability of ε with increasing X (i.e., nonconstant variance), then the range of residual deviations about zero will increase with increasing X. If the pattern of residuals describes a curved form, it is likely that the simple linear model is inadequate, and a higher order model should probably be tried (see Zar, p. 223).

It is often useful to make a histogram of the residuals (see Section 2) in order to see if the residuals appear to be normally distributed, or if there is marked skewness. The regression procedures in PROPHET enable you to make tables of residuals after they perform the initial analysis.

Other forms of residual plots are often useful. For example, the residuals may be plotted against $\hat{Y}_i$, to see if there is any systematic variation in the residuals.

H. Outliers

Sometimes a graph of residuals reveals certain points that appear to be clearly anomalous, in that they lie far away from all the other points (i.e., the absolute value of the residual is very large). Such points are known as outliers, and they should be carefully examined to determine if there is some logical justification for removing them from subsequent analysis. If the reason for an outlier can be traced to a known cause, such as failure in recording apparatus, or incorrect entering of the number, it may be excluded. You should always be careful, however, to ensure that the outlier is truly an invalid point rather than a datum that provides useful, albeit unusual, information.

I. Confidence curves about the regression line

The regression equation allows one to obtain an estimate of the true population value, Y_i, at a given value of X_i. The standard error of this predicted value is

$$SE(Y_i) = \sqrt{s^2_{y\cdot x}\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right)}$$

This standard error is a function of X, and has a hyperbolic shape with a minimum value at $X_i = \bar{X}$. When the function is multiplied by the t-value with degrees of freedom corresponding to the residual degrees of freedom and $\alpha=0.05$, these values may be plotted as positive and negative deviations about the regression line. The resultant lines are known as 95% confidence curves. (Note that if t corresponds to $\alpha=0.01$, the lines would be called 99% confidence curves.) An example of confidence curves superimposed on the regression line is shown in Figure 6-5.

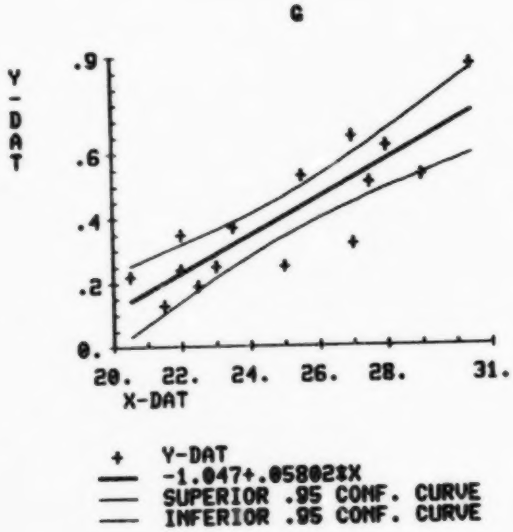


Figure 6-5

If 95% intervals are chosen, the interpretation is that the intervals will enclose the true mean of Y for a given X 95% of the time. Note that you should expect to see observed data points outside these confidence curves, since they are confidence curves of the mean of Y_i and not of a particular Y_i. A rough rule of thumb is that points which lie more than 3 standard deviations from the regression line may be outliers.

J. Weighted least squares

It may happen in regression analysis that some observations are not as reliable or accurate as others. This generally implies that the variances of any group of Y observations corresponding to a certain X will not all be equal, violating one of the basic assumptions of regression methodology. When you find this to occur, you can "weight" each data point by the inverse of its variance. In this way points that are not well determined (large variance) will not have such a large effect on the fit as do points with small variance. In PROPHET you can specify weights or variances in the regression analysis. Weights are assumed to be the inverse of the variance. That is if s_i^2 is the variance for Y_i and w_i is the weight then $w_i=1/s_i^2$. Tables 6-1 and 6-2 show the equations used in regression analysis. PROPHET standardizes the weights so that their sum equals the number of data points, thereby making various fits comparable. For a more detailed discussion of weighted regression, see Draper and Smith, Applied Regression Analysis or Dunn and Clark, Applied Statistics: Analysis of Variance and Regression.

K. Regression through the origin

Sometimes, for some theoretical reasons, you will know in advance that the intercept must be zero in order for the model to make sense. The hypothesis tests for regression lines through the origin are performed in the manner described above. Table 6-1 shows the equations used in this case. It should be noted that the regression line may not pass through the point $(\bar{X},\bar{Y})$. After the regression analysis has been performed with the PROPHET command FIT LINE, you are asked if you also want a line forced through the origin.

6.5 PROPHET Usage: FIT LINE and CONFCURVS

To fit a straight line to a curve of data points in a graph, use the following command:

```
FIT LINE TO [CURVE curve-number OF] graph
```

Data from a Graph

If the curve specification is omitted, curve 1 is used. The selected curve must have two or more data points.

FIT LINE produces the output shown in Figure 6-4. This display includes general information first: the number of data points, the correlation coefficient R, R-squared, and the standard error of regression. This is followed by two tables: the parameter table and the analysis of variance table.

The parameter table contains the computed value of the intercept and the slope of the fitted line, the standard error of the parameters, and results of the t-tests to test whether the parameters are zero.

The analysis of variance table contains the regression and residual sum of squares, degrees of freedom and mean square, along with the associated F-test.

To help you determine the adequacy of the fit you are then asked in sequence whether you would like to see:

- the line added to your graph.
- a table of residuals which includes the X and Y data, the fitted Y values and the residuals.
- a graph of the residuals vs. X.
- a graph of the residuals vs. the fitted Y.

If you answer "YES" to any of the questions, you will be shown the graph or table before the next question is asked.

Finally, FIT LINE asks whether you would like to repeat the fit forcing the line through the origin (the line will not contain an intercept term). The full analysis as described above will be repeated. Although mathematically it is possible to force the line through the origin, it is almost always inappropriate.

Weighted Fit

If each observation is not made with the same precision, you can perform a weighted fit using the following command:

```
FIT LINE TO [CURVE curve-number OF]  graph

                 |WITH WEIGHTS   IN COLUMN wt-col|
                 |     VARIANCES                 |
```

For example,

```
FIT LINE TO G WITH WEIGHTS IN COLUMN 2
```

The column of weights or variances must be added to the points table of your graph before performing the fit. You can do this with the command

```
ADD COLUMN TO POINTS OF graph
```

If the option WEIGHTS is used, then the values will be applied as a multiplier to the squared error term at each point. All weights must be positive values. If VARIANCES are used, the reciprocal of each variance is used as the weighting factor for each data point. The weights or inverse variances are adjusted to sum to n, the number of data points. In this way the magnitude of the residual sum of squares is comparable to the magnitude that would be obtained if no weights were used. There is no effect on any of the other statistics since only relative weighting is used.

Using Data from a Table

It is also possible to perform the analysis on data in a table without having to make a graph beforehand. The form of the command is

```
FIT LINE TO [COLUMN x-col VS COLUMN y-col] OF table

          |WITH WEIGHTS   IN COLUMN wt-col|
          |     VARIANCES                 |
```

For example,

```
FIT LINE TO COLUMN 5 VS COLUMN 2 OF TAB
```

If the column specifications for X and Y are ommited columns 1 and 2 are assumed, respectively. The presentation of the results of the analysis is identical to that described above. You are then asked whether you would like to see a graph of the fit, since you have not supplied a graph to add the fitted line to. If you respond "YES" a graph of data points and the fitted line is displayed.

CONFCURVS

You can add confidence curves to a line added to a graph by FIT LINE, using the public procedure CONFCURVS. To use this procedure, you type

```
RAC PUBLIC CONFCURVS
CALL CONFCURVS
```

The procedure will quiz you for the name of the graph to which confidence curves are to be added. If there are only two curves on the graph, it is assumed that curve 1 contains the data points and curve 2 the fitted line. If there are more than 2 curves you are asked which curve contains the data and which curve the fitted line. You are also asked to specify the confidence level, a FLOAT number between 0. and 1. (i.e., .95 for 95% confidence intervals). An example of the use of CONFCURVS is shown in Figure 6-6. The resulting graph is shown in Figure 6-5.

```
NAME OF GRAPH FOR WHICH YOU WANT CONFIDENCE CURVES ADDED ?G<

PLEASE SPECIFY CONFIDENCE LEVEL:?.95<
EQUATION OF SUPERIOR .95 CONFIDENCE CURVE IS
-1.046531+.05801858*X+2.16*SQRT(.01235684*(.06666667+ (X - 24.96667)**2/
**132.7333))

EQUATION OF INFERIOR .95 CONFIDENCE CURVE IS
-1.046531+.05801858*X-2.16*SQRT(.01235684*(.06666667+ (X - 24.96667)**2/
**132.7333))

HIT <GO> TO PROCEED?
```

Figure 6-6

6.6 Multiple Linear Regression

Sections 6.3 to 6.5 discussed regression techniques dealing with two variables, with one variable assumed to be dependent on the other. More complex models describe situations where the dependent variable Y is a function of more than one independent variable. For example, the increase in blood pressure in response to the infusion of a vasoactive drug may be a function of dose level and initial blood pressure. In general, the model may be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

The β's are called partial regression coefficients; each describes the dependence of Y on a particular X. Note that if p is the number of independent variables, at least $p+2$ data points are required to perform the multiple regression analysis.

The concept of least squares as applied to multiple regression can best be visualized by considering the case of two independent variables (X_1, X_2). If X_1 and X_2 are plotted as two orthogonal axes in the horizontal plane, and the dependent variable Y is plotted in the vertical direction, then the resultant equation describes a plane or surface, such that the sum of squared vertical deviations of observed Y points about this plane is a minimum. For more than two independent variables, the geometric interpretation becomes more complicated, but the concept is the same.

The assumptions for multiple regression are also similar to those for simple linear regression. That is, for each set of X-values (X_1, X_2 ...X_p) there exists a population of Y-values which are normally distributed with equal variance. The means of the true Y populations lie on the regression plane. The sums of squares are also partitioned into "regression" and "residual" components. The mathematics involved in multiple regression are rather complicated (refer to Dunn and Clark or Draper and Smith for a more detailed exposition), but interpreting standard errors, F-value, etc. is a simple extension of the single X-variable case.

In a multiple linear regression certain partial regression coefficients may turn out to be not significantly different from zero (i.e., when you examine their t-statistics). This may imply that the dependent variable Y is not significantly related to those particular X's. Alternatively, it may mean that the other X's in the equation explain the same variation in the Y's as the nonsignificant X's; since the value of any given coefficient is dependent on what other X-values are included in the analysis.

The determination of the final "best" model, that is, the decision of which X's to include in the final equation, is not a trivial one, and the criterion for what constitutes the "best" model is not always clear.

If there are p X-variables, there are 2^p regression lines that may be obtained, if we allow each set the option of either being included or not included in the model.

One approach to this problem is the so-called "stepwise regression procedure". Independent variables are added to or removed from the equation depending on their contribution to the overall fit of the model. The public procedure MULREG allows you to add and remove variables interactively.

6.7 PROPHET Usage: MULREG

The public procedure MULREG carries out an interactive version of a stepwise multiple regression analysis on either up to 100 variables, each with a maximum of 100 observations, or up to 19 variables, each with a maximum of 499 observations.

MULREG invokes a FORTRAN procedure to perform the regression analysis but the basic input data are assumed to be in a PROPHET table. However, because of the large amount of time required for table-to-file conversions, the program has been designed to create a permanent file of the table data so that all subsequent analyses on this data set are performed on the file rather than the table. Any number of data files can be created and saved for future analysis.

MULREG is called from PROPHET by the command

```
RAC PUBLIC MULREG
CALL MULREG
```

After you call MULREG, you will be prompted for commands to select from a number of options that have been built into the PROPHET portion of the program. These commands are:

W = Write the data array from the table into a file (you will be asked to supply a file name). After typing this command you will be asked to select one of the following subcommands: ST, ED, EX or W. ST is used to scan the table, ED allows you to edit the data (no changes are made in the original table), EX exits from the program, and W writes the data into the file. The regression program cannot handle empty data points, and any such cells will be replaced with 1.E-35 in the output file. Note that <u>only</u> columns of type "NUMBER" are allowed. Note also that if row names or column headings contain carriage returns, only the characters preceding the first return will be retained.

R = Run the Fortran stepwise regression program (you will be asked to supply the file name containing the data). A description of this program is below. When the program returns to PROPHET you can add a column of fitted Y-values to the table using the final regression equation (this is done as a derived column).

L = List directory of data files. Whenever a file is created using the "W" command, an entry is made in a directory file called "TABFIL.DIR" of the table name, file name, number of rows and columns, and the date and time of creation.

DF = Delete a file from the directory. You use this command to delete any file, not only the data files created by the "W" command.

SF = Scan a file. You use this command to scan any file. Its major purpose is to look at the output files from the regression analysis (see below).

ST = Scan the table.

EX = Exit; return to PROPHET command level.

To speed up the printout of the regression data, you can specify (when asked) a short printout that consists of only the variables in the regression and the first few variables not in the regression. The program produces a copy of the correlation matrix and a copy of the regression analysis results. These two files can be scanned with the "SF" command described above. At each step in the regression analysis, you are given the option of viewing the residuals. In this way you can eliminate badly deviating data points before continuing with the analysis. Since the files containing the correlation matrix and the regression analysis results can be quite big, you should delete these files (e.g., by using the "DF" command above) as soon as they are no longer of use. Note that very little error checking is done in the Fortran program so you should exercise due care in entering your responses. All commands in the FORTRAN program must be ended with the return key (not <GO>). Figure 6-7 illustrates how you might use MULREG.

```
COMMAND (TYPE 'H' FOR HELP) ?  H<
 W  = WRITE (EDITED) TABLE ONTO FILE
 L  = LIST FILE DIRECTORY
 R  = RUN STEPWISE REGRESSION
 SF = SCAN A FILE
 ST = SCAN TABLE
 DF = DELETE FILE FROM DIRECTORY
 EX = EXIT
COMMAND (TYPE 'H' FOR HELP) ?


SUB COMMAND (TYPE 'H' FOR HELP) ?  H<
 ST = SCAN TABLE
 ED = EDIT TABLE (DELETE ROWS AND COLS)
 EX = EXIT
 W  = WRITE TABLE ONTO FILE
SUB COMMAND (TYPE 'H' FOR HELP) ?


COMMAND (TYPE 'H' FOR HELP) ?  R<
FILE NAME ?  MMM<   YOUR FILE HAS BEEN TEMPORARILY RENAMED TO
'STPRG.DAT'. IF NORMAL EXECUTION OCCURS, THE ORIGINAL FILE NAME
WILL BE RESTORED AT THE END OF THE ANALYSIS. SHOULD THE COMPUTER
CRASH OR THE FORTRAN PROGRAM FAIL, YOUR DATA WILL BE LEFT ON THE
FILE 'STPRG.DAT' AND YOU WILL HAVE TO RESTORE THE ORIGINAL
FILE NAME BY USING THE 'RENAME FILE' COMMAND IN PROPHET

FORTRAN PROGRAM: ENTER ALL DATA WITH THE RETURN KEY

         DO YOU WISH TO PRINT THE ROW NAMES ?
         N


         NO. ROWS TO BE DELETED ?
         0


         COL NAMES

           1  DEG
           2  CM
           3  MM
           4  MIN
           5  ML


         NO. COLS TO BE DELETED ?
         0



         DEPENDENT VARIABLE ?
         5
```

Figure 6-7a

```
VARIABLE X ( 1) ENTERED AT F=  35.535

MULTIPLE CORRELATION COEFFICIENT        R= 0.73081
  CORRECTED FOR DEGREE OF FREEDOM    R(C)= 0.72045
STANDARD ERROR OF ESTIMATE              = 0.47079
DEGREES OF FREEDOM                      = 31

VARIABLE                  CONTRIB.    REGRESSION   STANDARD   T TEST OF
                           TO RSQ     COEFF. (B)  ERROR OF B   B COEFF.

 1 DEG                 *  -0.53408   -.1292      .2167E-01 -5.961
 4 MIN                     0.12326
 3 MM                      0.00464
 2 CM                      0.00089
 5 ML                    DEPENDENT VARIABLE
* VARIABLE IN EQUATION

CONSTANT =        3.0496
OVERALL F=  35.535


DO YOU WISH TO SEE THE RESIDUALS ?
N


PLEASE SAY WHETHER TO ENTER A VARIABLE I (+I), TO REMOVE IT (-I),
 OR TO EXIT (0).  STEPWISE SOLUTION SUGGESTS   4
4

VARIABLE X ( 4) ENTERED AT F=  10.791

MULTIPLE CORRELATION COEFFICIENT        R= 0.81076
  CORRECTED FOR DEGREE OF FREEDOM    R(C)= 0.79655
STANDARD ERROR OF ESTIMATE              = 0.41042
DEGREES OF FREEDOM                      = 30

VARIABLE                  CONTRIB.    REGRESSION   STANDARD   T TEST OF
                           TO RSQ     COEFF. (B)  ERROR OF B   B COEFF.

 1 DEG                 *  -0.55958   -.1324      .1891E-01 -6.999
 4 MIN                 *  -0.12326    .2013      .6129E-01  3.285
 2 CM                      0.00099
 3 MM                      0.00023
 5 ML                    DEPENDENT VARIABLE
* VARIABLE IN EQUATION

CONSTANT =        2.5520
OVERALL F=  28.775


DO YOU WISH TO SEE THE RESIDUALS ?
N


PLEASE SAY WHETHER TO ENTER A VARIABLE I (+I), TO REMOVE IT (-I),
 OR TO EXIT (0).  STEPWISE SOLUTION SUGGESTS   0
0


PROPHET: ENTER ALL DATA WITH THE GO KEY


YOUR DATA FILE HAS BEEN RESTORED TO ITS ORIGINAL NAME.

DO YOU WISH TO DELETE THE CORRELATION
MATRIX FILE ?Y
```

Figure 6-7b

6.8 Polynomial Regression

Polynomial regression involves fitting a set of Y-observations to a polynomial expansion in X. That is,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$$

In certain situations there are theoretical reasons for believing the polynomial to be correct. It should be remembered, however, that indiscriminately allowing the order (i.e., the exponent of the highest power of X) to increase so as to obtain better fits is a totally spurious approach. In fact, you can fit any set of n data points exactly by choosing the order of the polynomial to be equal to n-1.

As in the other forms of regression we have discussed, polynomial regression uses the least squares approach, minimizing the sum of squared vertical deviations of individual Y's about their corresponding predicted values.

Mathematically, polynomial regression is an example of the multiple regression methods discussed in Section 6-5. The approach is to reassign the distinct X-variables in multiple regression such that X_1 becomes X, X_2 becomes X^2, X_3 becomes X^3, etc. Then the models are mathematically equivalent, and the analysis of variance is performed in an identical fashion.

The F-ratio, standard error of the estimate, t-values for partial regression coefficients, and so on, are also to be interpreted as in linear regression.

Note that the PROPHET command FIT POLYNOMIAL (see below) will always produce a polynomial consisting of all powers of the independent variable X, up to the maximum order specified. Sometimes you may know that the model you are trying to fit should consist of only specific powers of X, say $Y = \beta_0 + \beta_1 X^2 + \beta_2 X^4$. Specific functions of this sort may be analyzed using the MULREG procedure described in Section 6.7. To do this with MULREG you would create a derived column (called the X_1 variable in MULREG) which would be set equal to X^2, and similarly, create a second derived column (called X_2), set equal to X^4. Then you call MULREG to perform the analysis. You can also use FIT FUNCTION to perform the analysis.

6.9 PROPHET Usage: FIT POLYNOMIAL

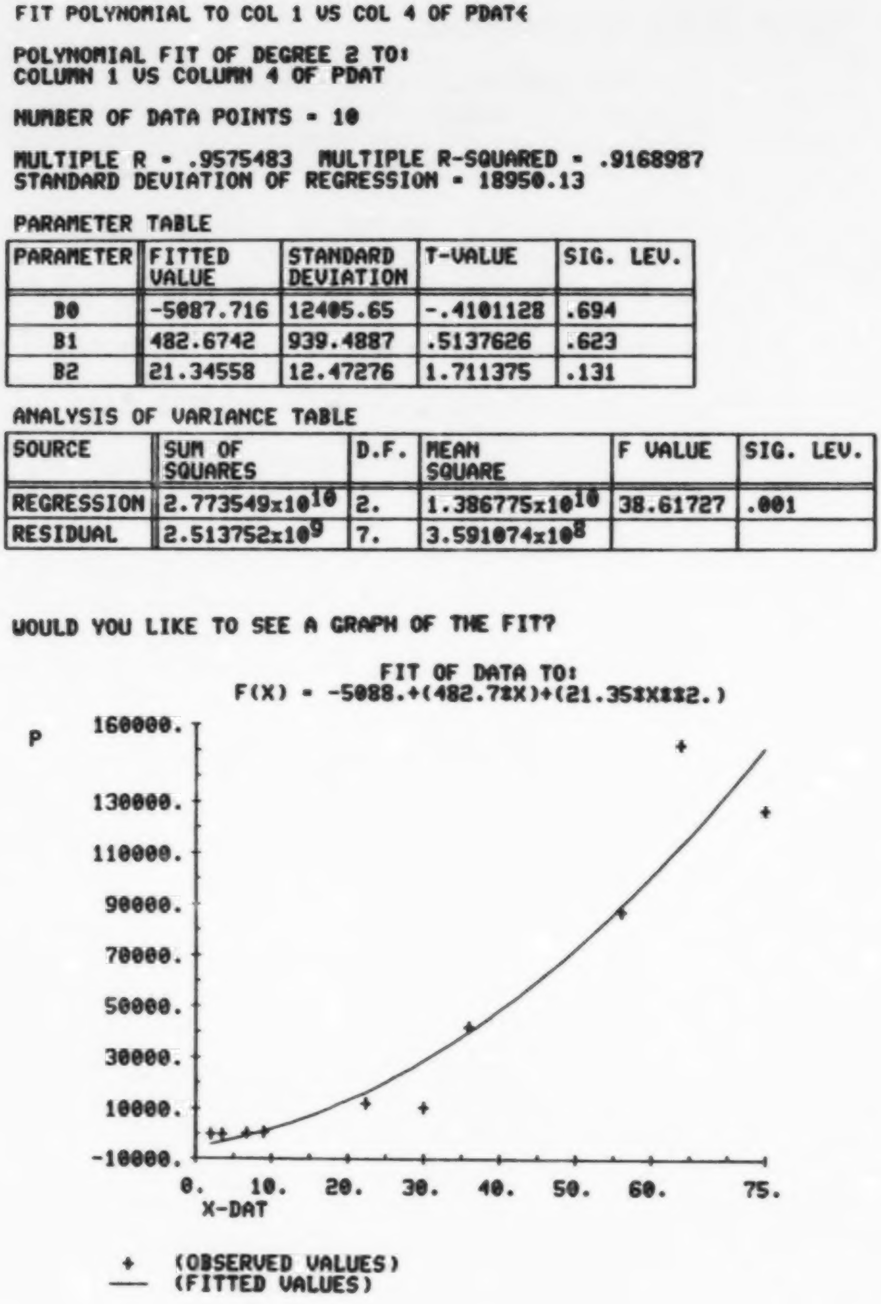
The command to fit a polynomial to a set of data in a table or graph is identical to the FIT LINE command, except that the word POLYNOMIAL is used in place of LINE. For example, the following are valid commands:

```
FIT POLYNOMIAL TO TABDATA WITH WEIGHTS IN COLUMN 3
FIT POLYNOMIAL TO CURVE 2 OF POLGRAPH
FIT POLYNOMIAL TO COLUMN 4 VS COLUMN 2 OF TAB
```

First you will be asked what degree polynomial you wish to fit to your data. You must type in an integer number greater than zero. As in FIT LINE, the summary of statistics and results will be reported. The coefficients of the polynomial are labelled B0, B1, B2, etc., so that B0 is the intercept, B1 the coefficient for X, B2 the coefficient for X^2, etc. The results of the FIT POLYNOMIAL command are shown in Figure 6-8.

As in FIT LINE, after the results are displayed you are given the options of seeing a graph of the fitted curve, a table of residuals, a graph of residuals vs. X, and a graph of the residual vs. Y.

Finally, you are asked if you would like to fit a polynomial of the next higher order. If you respond "NO" the analysis will stop. But if you respond "YES" the whole analysis will be repeated with the higher order polynomial.



FIT POLYNOMIAL TO COL 1 VS COL 4 OF PDAT<

POLYNOMIAL FIT OF DEGREE 2 TO:
COLUMN 1 VS COLUMN 4 OF PDAT

NUMBER OF DATA POINTS = 10

MULTIPLE R = .9575483 MULTIPLE R-SQUARED = .9168987
STANDARD DEVIATION OF REGRESSION = 18950.13

PARAMETER TABLE

PARAMETER	FITTED VALUE	STANDARD DEVIATION	T-VALUE	SIG. LEV.
B0	-5087.716	12405.65	-.4101128	.694
B1	482.6742	939.4887	.5137626	.623
B2	21.34558	12.47276	1.711375	.131

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F VALUE	SIG. LEV.
REGRESSION	2.773549×10^{10}	2.	1.386775×10^{10}	38.61727	.001
RESIDUAL	2.513752×10^{9}	7.	3.591074×10^{8}		

WOULD YOU LIKE TO SEE A GRAPH OF THE FIT?

FIT OF DATA TO:
F(X) = -5088.+(482.7*X)+(21.35*X**2.)

P
160000.
130000.
110000.
90000.
70000.
50000.
30000.
10000.
-10000.
0.
10.
20.
30.
40.
50.
60.
75.
X-DAT

+ (OBSERVED VALUES)
— (FITTED VALUES)

Figure 6-8

Improvement in the Fit

If the higher order fit is performed, in addition to the parameter and analysis of variance tables, information about the improvement in the fit will be displayed. The improvement in multiple R-squared is the difference between the R-squared in the present fit and the previous fit. This will always be a positive number. The improvement in the standard deviation of regression is the difference in that statistic between the previous fit and the present fit. If there was an improvement the value will be positive; otherwise, it will be negative. The partial F-value is equivalent to the t-value for the parameter of highest degree; it tests whether there is a significant improvement in the fit by adding the additional degree. If it is not significant, you should probably stop the analysis and reconsider the previous fit.

6.10 PROPHET Usage: FIT FUNCTION

You can fit any linear or nonlinear function to data in a graph or table using the FIT FUNCTION command. The command format is very similar to those for FIT LINE and FIT POLYNOMIAL. For example:

```
FIT FUNCTION TO DATATAB
FIT FUNCTION TO CURVE 2 OF DATAGRAPH
```

In FIT FUNCTION, the fitting is done in a series of trials, or iterations, in which the parameters are systematically adjusted until a least-squares solution is reached. The refinement of the parameters is done by MLAB using a Marquardt-Levenberg method. This technique combines the advantages of the Gauss-Newton method and the steepest descent approach. For a complete description of MLAB see <u>MLAB</u>: <u>An On-Line Modelling Laboratory</u>, Division of Computer Research and Technology, National Institutes of Health.

Convergence

MLAB decides whether or not the fitting algorithm has converged based on the reduction in the residual sum of squares error from one iteration to the next. If the ratio of the change in the residual sum of squares to the current residual sum of squares is greater than the convergence factor (normally set at .001), then the fit is considered to be still improving and therefore not converged. The fit has converged when the algorithm cannot improve the parameter values any further with the data given. Convergence of the algorithm therefore has no relation to goodness of fit of the fitted function, which should be determined by the standard errors of the regression and coefficients.

Form of Function

When you invoke FIT FUNCTION, you will first be asked to supply the form of the function as a PROPHET expression. The expression is a function of a single independent variable X and can include any number of parameters and constants. Any of the mathematical functions listed in the COMMAND SUMMARY can be used. The expression <u>cannot</u> contain calls to user-defined procedures.

The parameters of the function must be named "B1" through "Bn" and must first occur in lexicographic order when reading the function from left to right. Thus:

```
1/(B1*X)
B1 * SIN (B2*X) + B3
B1*X**B2/(X**B2 + B3**B2)
```

are legal functions, whereas

```
B1 * SIN (B3*X)
B3 * EXP(B2*X) + B1
```

are not.

You can also supply the function expression directly in the command. For example

```
FIT FUNCTION 'B1 + B2 * EXP(B3*X)' TO DATATAB
FIT FUNCTION F TO COLUMN 4 VS COLUMN 5 OF DATATAB
```

where F is a TEXT variable containing the expression of the function.

Initial Estimates

After the form of the function has been supplied you will be asked whether you would like to supply initial estimates for the parameters. Because an iterative procedure is used to systematically adjust the parameters, supplying initial estimates can greatly speed the process, avoid widely functuating adjustments to the parameters, and more importantly, incease the chance of finding the best fit rather than some local minimum that is not the best.

You might try graphing a function along with your data to arrive at a guess. Even having the correct sign, if not the magnitude, for the initial estimate of a parameter can be helpful. If you supply initial guesses you will be prompted for their values. Each must be a FLOAT number ended with a <GO>. If you do not supply an initial guess, the value 1.0 will automatically be used.

Constraints

Next you are asked whether you would like to enforce some constraints on the parameters. You may know, for instance, that the value of a parameter must be positive because of theoretical considerations. This can be useful for the fitting process by eliminating some areas to search over. Usually, however, an initial guess will serve this purpose better.

Constraints are specified as a linear relationship between a parameter (or combination of parameters) and a certain constant value. The relationship is denoted using the following symbols:

<	less than
>	greater than
=	equal to
<=	less than or equal to
>=	greater than or equal to

For example, B1 < 14.0 and B2 >= 0 are valid constraints. Parameters may also be combined, such as B2 + B3 < 100. You can also divide or multiply any parameter by a constant. For example, B1/4 + 2 * B3 > 10. If you want to enforce constraints you will be asked to type them one at a time, each followed by a <GO>. Another <GO> ends the list of constraints.

Constraints are considered "active" after the fitting process if equality holds in the constraint equation. For example, with the above constraints, if after the fitting process

```
        B1 = 14.0
  or    B2/4 + 2*B3 = 0.0
```

then each constraint is active. This means that the true minimum is outside the constraint region and hence the minimum within the constraint region is at the boundary. If there are active constraints, the succeeding statistical analysis will not be meaningful.

Output

After you specify any constraints PROPHET will type

```
[PERFORMING FIT ...
```

This indicates that MLAB has been invoked to perform the minimization. Do <u>not</u> try to abort from the FIT FUNCTION command until PROPHET has typed

```
DONE]
```

At that point you will be back in PROPHET.

When the fit is complete, you will then see the results of the parameter estimates and goodness of fit statistics, you will next be given the options to see a graph of the fit, a table of residuals and graphs of the residuals. An example of the use of FIT FUNCTION is shown in Figure 6-9.

```
ENTER EQUATION OF FUNCTION TO FIT TO DATA:
F(X) = ?B1 + B2/X<

WOULD YOU LIKE TO SPECIFY INITIAL VALUES FOR THE PARAMETERS?
(IF NOT, ALL PARAMETERS WILL HAVE INITIAL VALUES 1.0)?Y<
        INITIAL VALUE FOR B1?4.<
        INITIAL VALUE FOR B2?0.5<

WOULD YOU LIKE TO ENFORCE CONSTRAINTS?N< PERFORMING FIT...


FIT OF FUNCTION B1 + B2/X TO:
COLUMN 1 VS COLUMN 2 OF FF

NUMBER OF DATA POINTS = 10

MULTIPLE R-SQUARED = .7887954
STANDARD DEVIATION OF REGRESSION = .09503203
```

PARAMETER TABLE

PARAMETER	FITTED VALUE	STANDARD DEVIATION	T-VALUE	SIG. LEV.
B1	5.023119	.04497666	111.6828	.0001
B2	.6244975	.1142494	5.46609	.001

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F VALUE	SIG. LEV.
REGRESSION	.2698305	1.	.2698305	29.87796	.001
RESIDUAL	.0722487	8.	.009031088		

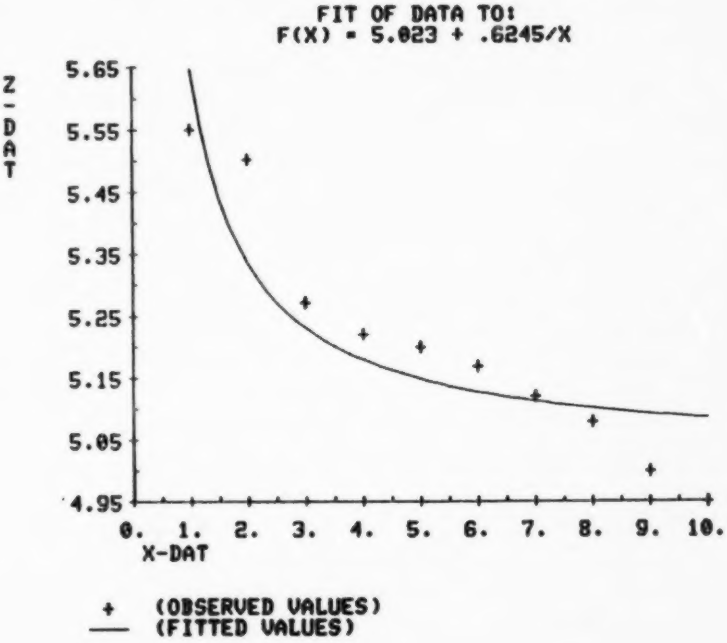


Figure 6-9

6.11 PROPHET Usage: FITFUN

The public procedure FITFUN, like FIT FUNCTION, fits any linear or nonlinear function to observed data by invoking MLAB. FITFUN offers more extensive capabilities: you can use functions containing more than one independent variable and fit several functions simultaneously. When a system of functions describes the model, the same parameters can be in more than one of the functions. The parameterized functions may include any combination of powers, logarithms, trigonometric functions and hyperbolic functions; for parameter names and independent variable names, you can use any legal PROPHET variable names.

Form of Function

Examples of legal functions in FITFUN are:

```
F1 = A*SIN(B*X1) + H*EXP(Q1*X2)
F2 = YMAX * A**P/(A**P + K**P)
```

Function and Parameter Tables

With FITFUN, you use a function table and a parameter table to represent the system you want fitted. When FITFUN is called you will be able to construct either of these tables if you do not already have them. The function table simply contains the name and equation of each function in the system you want to fit. The parameter table contains the name of each parameter in the system plus a corresponding initial value which will be used to initiate the fit. FITFUN provides the best fits when the initial parameter values are well specified. Assuming that your system is a good model of the observed data, your objective then should be to modify carefully the parameter specifications in order to get as good a fit as possible. The parameter table serves as a data base for comparing the fitted parameter values of various fits. Prior to performing a fit, you will be able to edit either table.

Output

Output from FITFUN is similar to that of the other FIT procedures; it includes a parameter table with the fitted parameter values, their standard errors and significance levels. For each function included in the system to fit, an analysis of variance table for the regression, the standard error of regression, the multiple correlation coefficient R^2 and other descriptive information will be supplied. Finally, you have the option of seeing a table of residuals and a graph of the fit for each function.

Many parameters

When the system of functions contains many unknown parameters, estimates of the parameters may fluctuate widely during the iterative fitting process. To control this you can either constrain some of the parameters when you know some of their properties, or you can temporarily fix some parameters to their initial values. The resulting fit may provide good initial estimates for the unconstrained model. After constraining several sets of parameters in a series of fits you can then repeat the fit with no constraints, using the best initial guesses.

Calling FITFUN

To use FITFUN, type the following commands:

```
RAC PUBLIC FITFUN
CALL FITFUN
```

In order to perform a fit, FITFUN will ask you for the following information:

- the names of those functions to include in the system to fit;
- which parameters to vary in the next fit (if a parameter does not vary, it will be set equal to its initial value throughout the fit);
- any constraints desired;
- the column of your parameter table with initial parameter values;
- the name of your data table;
- the columns of your data table with values for the independent variables;
- the columns of your data table with observed dependent values;
- any weights desired (you will be asked to specify the column of your data table that contains weights).

Results

Next, FITFUN performs the specified fit. Results of the fit include a parameter results table. For each varied parameter, this includes its fitted value, standard error, t-value and significance level.

For each <u>function</u> in the system to fit FITFUN displays:

- a summary of the fit (this includes other statistics in addition to an analysis of variance table);
- if you request it, a table of residuals; and
- if you want, a graph of the fit;

After all the results have been presented, you will be able to save the fitted parameter values in your parameter table. You may save these values in either a new column or an old one. If the fitted parameters replace an older set of values, only those parameters which varied will have their values updated. Moreover, the column heading will include the F-values of each function in the system. Thus, you will be able to compare the goodness of two fits by their column headings. In addition, the next column will be set to the standard errors of the fitted parameters.

You may then repeat a fit, subject to new parameter specifications (e.g., which parameters to vary or which initial parameter values to use in the next fit).

An example using FITFUN is shown in Figure 6-10.

```
 CALL FITFUN <

PLEASE SPECIFY THE FUNCTION NAME AND THE EQUATION OF THE
FUNCTION TO BE ADDED TO YOUR FUNCTION TABLE WHEN PROMPTED. WHEN DONE,
RESPOND WITH "(EX)IT" TO THE FUNCTION NAME PROMPT.

         FUNCTION NAME?F1<
         EQUATION OF F1?B1*SIN(X1) + B2*LOG(X2)<

         FUNCTION NAME?EXIT

WHEN PROMPTED, PLEASE SPECIFY THE PARAMETER NAME PLUS
ITS INITIAL VALUE FOR EACH PARAMETER YOU WISH TO SAVE IN THE
TABLE. WHEN DONE, RESPOND WITH "(EX)IT" TO THE PARAMETER NAME PROMPT.

         PARAMETER NAME?B1<
         INITIAL PARAMETER VALUE FOR B1?1.<

         PARAMETER NAME?B2<
         INITIAL PARAMETER VALUE FOR B2?1.<

         PARAMETER NAME?B3<
         INITIAL PARAMETER VALUE FOR B3?0.5<

         PARAMETER NAME?EXIT
```

Figure 6-10a

```
SYSTEM TO FIT CONSISTS OF:

F1 = B1*SIN(X1) + B2*LOG(X2)

LIST OF THOSE PARAMETERS (SEPARATED BY COMMAS OR
BLANKS) TO VARY IN THE NEXT FIT?B1 B2
←

WOULD YOU LIKE TO ENFORCE ANY CONSTRAINTS IN THE NEXT FIT?N←

NAME OF DATA TABLE?HP←

        COLUMN OF HP WITH VALUES FOR X1?1←

        COLUMN OF HP WITH VALUES FOR X2?2←

        COLUMN OF HP WITH EXPERIMENTAL VALUES FOR F1?3←

WOULD YOU LIKE TO USE WEIGHTS OR VARIANCES?N


FIT FOR FUNCTION F1 = B1*SIN(X1) + B2*LOG(X2) TO:
COLUMNS 1,2 VS COLUMN 3 OF HP
NUMBER OF DATA POINTS = 10
MULTIPLE R-SQUARED = .993504
STANDARD DEVIATION OF REGRESSION = .3372294

PARAMETER TABLE
+---------+---------+---------+---------+----------+
|PARAMETER|FITTED   |STANDARD |T-VALUE  |SIG. LEV. |
|         |VALUE    |DEVIATION|         |          |
+---------+---------+---------+---------+----------+
|   B1    |1.033145 |.1532299 |6.742448 |.001      |
|   B2    |1.169691 |.0359037 |32.57855 |.0001     |
+---------+---------+---------+---------+----------+

ANALYSIS OF VARIANCE TABLE
+----------+--------+----+--------+--------+---------+
|SOURCE    |SUM OF  |D.F.|MEAN    |F VALUE |SIG. LEV.|
|          |SQUARES |    |SQUARE  |        |         |
+----------+--------+----+--------+--------+---------+
|REGRESSION|139.1445|2.  |69.57226|611.7659|.0001    |
|RESIDUAL  |.9097894|8.  |.1137237|        |         |
+----------+--------+----+--------+--------+---------+
```

Figure 6-10b

FITFUNSAV When you want to save any piece of information in the results of the fit as a top-level variable, type

```
RAC PUBLIC FITFUN
CALL FITFUNSAV
```

Operation of this procedure is then identical to FITFUN.

FITFUNQK When you want to perform a "quick fit," without creating a paramter table or function table, you type

```
RAC PUBLIC FITFUN
CALL FITFUNQK
```

Unlike FITFUN, this invocation will perform only a single fit. You will be asked for all the information (see FITFUN) required for performing the fit.

After the fit is performed, the results will be presented in a fashion identical to that of FITFUN. The only difference is that you will not be able to repeat the fit.

FITFUNQKSAV

To perform a combination of FITFUNQK and FITFUNQKSAV you type

```
RAC PUBLIC FITFUN
CALL FUTFUNQKSAV
```

Thus you use this call when you want to perform a "quick fit", and to be able to save information in the results of the fit.

6.12 PROPHET Usage: EXPFIT

EXPFIT is a regression procedure which fits data to a function of the form

$$Y = \beta_0 + \beta_1 e^{\beta_2 X} + \ldots + \beta_{p-1} e^{\beta_p X}$$

One to four exponential terms may be included, either with or without the constant term. The procedure uses FIT FUNCTION to perform the actual fitting process and therefore requires initial estimates for all parameters.

Peeling

Estimates may be arrived at through an interactive "peeling" process. Peeling involves

- fitting a straight line to the tail of the log of the data;
- extrapolating the line back to X=0;
- subtracting it from the remaining points (not in the tail);
- fitting a new straight line to the next portion of the data;
- extrapolating this line, and so forth.

This process yields good initial estimates for the exponential fitting program itself, usually avoiding situations which cause the fitting algorithm either to take a very long time to converge or, more drastically, not to converge at all. After you have told the program the number of exponential terms to include, you then specify the range of X-values to be covered by this term in peeling. These ranges are asked for in reverse order, starting with the tail. After each stage of the peeling process is completed, you can still go back and change the previous estimates, and at each stage you have an opportunity to examine a plot of the fit calculated by the peeling.

Weighted Analysis

When you want to perform weighted analysis, you include the weights in a column of the data table. For exponential functions, weights often tend to markedly increase the chances of getting a good fit.

Calling EXPFIT

To use EXPFIT you type

```
RAC PUBLIC EXPFIT
CALL EXPFIT(table)
```

You then specify which columns contain the X and Y data, whether you wish to use a weighting factor, whether you would like a semi-log plot of the data, the number of exponential terms, the constant term (if any) and whether you want to perform peeling. An example using this procedure is shown in Figure 6-11a-d.

```
CALL EXPFIT(ETAB) €


COLUMN NUMBER FOR X VALUES =  1€

COLUMN NUMBER FOR Y VALUES =  2€

DO YOU WISH TO ASSIGN WEIGHTING FACTORS TO THE DATA?Y€
COLUMN NUMBER FOR WEIGHTS =  3€

DO YOU WISH TO SEE A SEMI-LOG PLOT OF THE DATA?Y
```

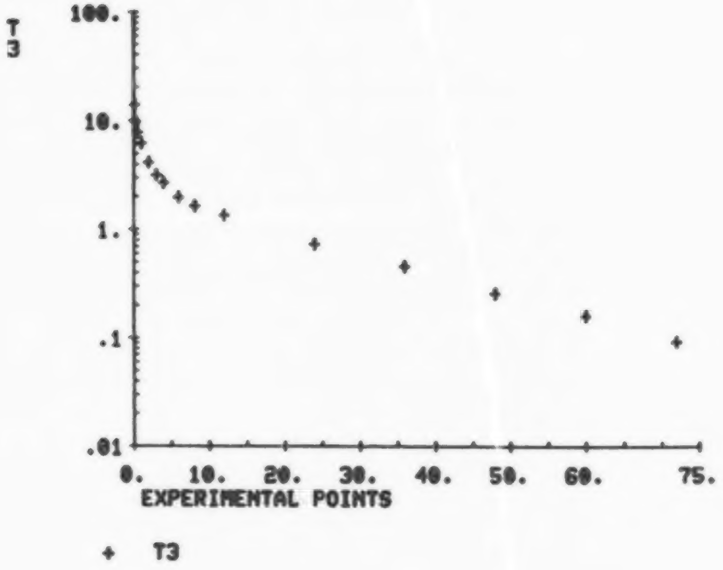
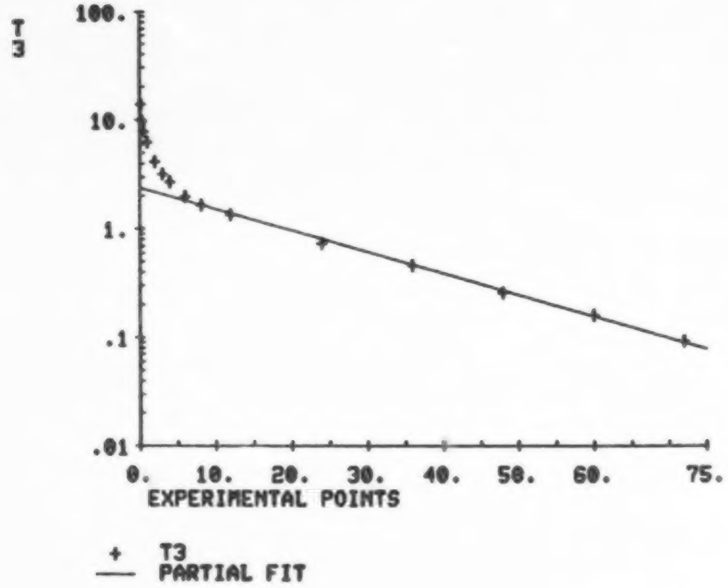


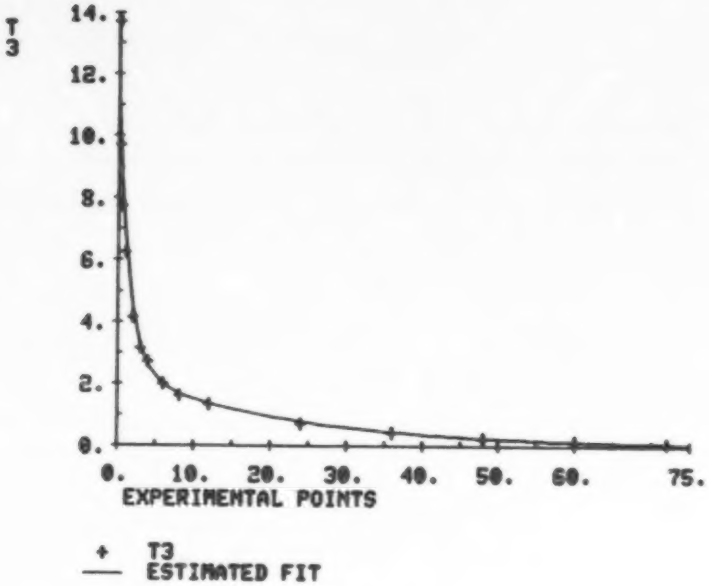
Figure 6-11a

```
FOR COMPONENT 2 : X >  6€  AND X <  75€

COEFFICIENT = 2.371862    EXPONENT = -.04560077

DO YOU WISH TO SEE A GRAPH OF THE FIT WITH THE PEELING DONE THUS FAR?Y
```



```
FOR COMPONENT 1 : X >  .1€  AND X <  5.€

COEFFICIENT = 8.87011    EXPONENT = -.6682957

DO YOU WISH TO CHANGE CURRENT ESTIMATES AND ENTER YOUR OWN GUESSES?N€

DO YOU WISH TO SEE A PLOT OF THE INITIALLY
ESTIMATED CURVE WITH THE OBSERVED POINTS?Y€
LINEAR OR LOG PLOT ('LIN' OR 'LOG')? LIN€
```

Figure 6-11b



```
(WEIGHTED) FIT OF FUNCTION B1*EXP(B2*X)+B3*EXP(B4*X) TO:
COLUMN 1 VS COLUMN 2 OF EXPDATA WITH WEIGHTS IN COLUMN 3

NUMBER OF DATA POINTS = 15

MULTIPLE R-SQUARED = .9940521
STANDARD DEVIATION OF REGRESSION = .02581872
```

PARAMETER TABLE

PARAMETER	FITTED VALUE	STANDARD DEVIATION	T-VALUE	SIG. LEV.
B1	9.146537	.7143765	12.80353	.0001
B2	-.7186727	.07935596	-9.056316	.0001
B3	2.333019	.144226	16.17613	.0001
B4	-.04534206	.001482499	-30.58489	.0001

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F VALUE	SIG. LEV.
REGRESSION	1.225475	4.	.3063688	459.5947	.0001
RESIDUAL	.007332671	11.	.0006666064		

Figure 6-11c

```
WOULD YOU LIKE TO SEE A GRAPH OF THE FIT?Y<


DO YOU WISH TO HAVE A TOP-LEVEL COPY OF THIS GRAPH?N<
```

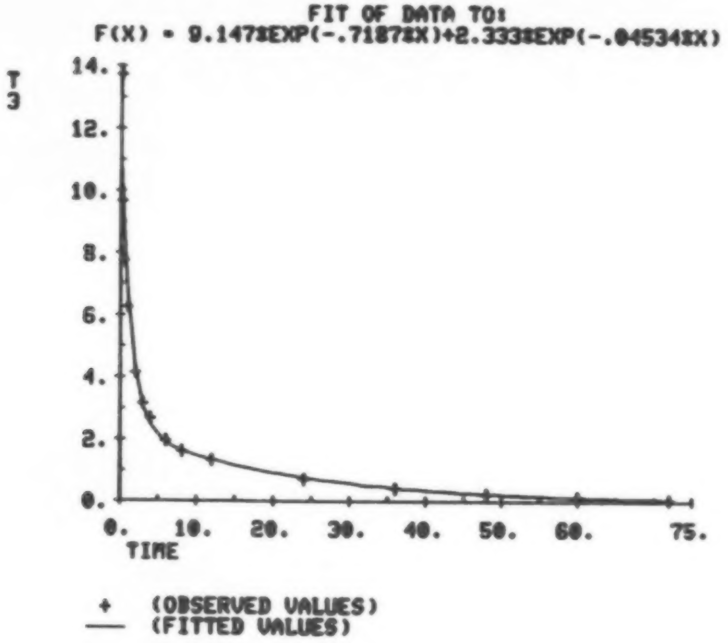


Figure 6-11d

6.13 PROPHET Usage: HYPERBOLIC

The public procedure HYPERBOLIC fits data to the enzyme kinetic equation

$$v = \frac{V_{max} [S]}{[S] + K_m}$$

where v = initial velocity of enzyme reaction
Vmax = maximum velocity
[S] = substrate concentration
K_m = Michaelis constant

The fit is carried out iteratively using a Taylor expansion of the nonlinear parameter Km.

The analysis is carried out in two steps:

. First, an initial estimate of Km, necessary for the iterative fit, is determined by carrying out a weighted linear regression of the double-reciprocal form of the equation

$$\frac{1}{v} = \frac{K_m}{V_{max}} \cdot \frac{1}{[S]} + \frac{1}{V_{max}}$$

. Second, using the initial estimate of Km calculated in the first step, fitting of the hyperbolic equation is continued until a convergence in the value of Km is obtained.

To use the procedure, you type

```
RAC PUBLIC HYPERBOLIC
CALL HYPERBOLIC(table)
```

The first column of the table should contain the X-values (substrate concentrations) and the second column should contain the Y-values (initial velocities). If a weighted fit is desired, the variances of the Y-data should be in column 3 of the table. (Note that the reciprocals of the variances are used as weighting factors in performing the least squares fit.) If column 3 contains any EMPTY cells, the weights for the corresponding points will be the average of the other weights.

HYPERBOLIC will print the values of Km and Vmax, and their standard errors. Then a graph of the data and fitted hyperbolic curve will be displayed. In addition a graph containing the double reciprocal (Lineweaver-Burke) plot will be displayed.

An example of the use of HYPERBOLIC is shown in Figure 6-12.

```
CALL HYPERBOLIC(HDAT)

   K = 178.9115           STD. ERR. K = 6.004304
VMAX = 3.690378        STD. ERR. VMAX = .07168832

    CONTINUE  ?Y←

   K = 178.9151           STD. ERR. K = 6.020699
VMAX = 3.690417        STD. ERR. VMAX = .07180676

    CONTINUE  ?Y←

   K = 178.9152           STD. ERR. K = 6.020914
VMAX = 3.690418        STD. ERR. VMAX = .07180834

    CONTINUE  ?Y←

   K = 178.9153           STD. ERR. K = 6.020918
VMAX = 3.690419        STD. ERR. VMAX = .07180836
```

Figure 6-12a

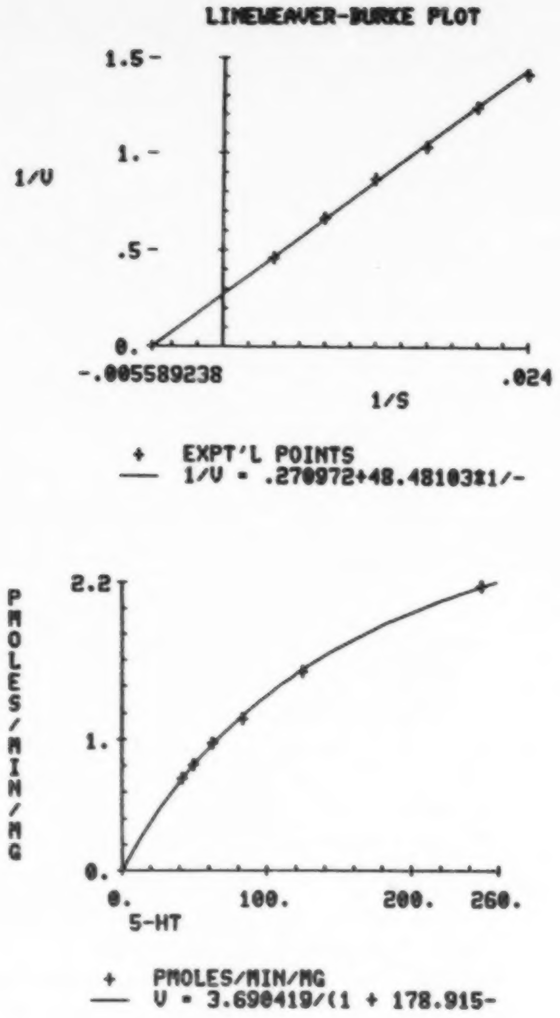

LINEWEAVER-BURKE PLOT
1.5
1.
.5
0.
1/V
-.005589238
.024
1/S
+ EXPT'L POINTS
— 1/V = .270972+48.48103*1/-
2.2
1.
0.
PMOLES/MIN/MG
0.
100.
200.
260.
5-HT
+ PMOLES/MIN/MG
— V = 3.690419/(1 + 178.915-

Figure 6-12b

TABLE 6-1. COMPUTING THE ANALYSIS OF VARIANCE IN REGRESSION.

Source	Model with Constant Parameter		Model with no Constant Parameter	
	Sum of Squares	d.f.	Sum of Squares	d.f.
Regression	$\sum_{i=1}^{n} w_i (\hat{Y}_i - \overline{Y})^2$	p − 1	$\sum_{i=1}^{n} w_i \hat{Y}_i^2$	p
Residual	$\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2$	n − p	$\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2$	n − p
Total	$\sum_{i=1}^{n} w_i (Y_i - \overline{Y})^2$	n − 1	$\sum_{i=1}^{n} w_i Y_i^2$	n

n – number of observations

p – number of parameters (including constant parameter)

$w_i = 1$ for unweighted regression

$\sum_{i=1}^{n} w_i = n$

$\overline{Y} = \sum_{i=1}^{n} w_i Y_i / \sum_{i=1}^{n} w_i$

$w_i \propto 1/s_i^2$, where s_i^2 is the variance of Y_i.

TABLE 6-2. BASIC REGRESSION TERMS.

Total SS = Regression SS + Residual SS

Mean Square = Sum of Squares/Degrees of Freedom

Standard Error of Regression = $\sqrt{\text{Residual MS}}$

$$F = \frac{\text{Regression MS}}{\text{Residual MS}}$$

$$\text{Multiple } R^2 = \frac{\text{Regression SS}}{\text{Total SS}}$$

$$\text{t-value} = \frac{\text{parameter}}{\text{SE of parameter}}$$

7. ANVAR

SECTION 7

MULTIFACTOR ANALYSIS OF VARIANCE

7.1 Introduction

In many situations there may be more than one factor influencing each observation. For example, the blood presure response of different strains of mice to several dose levels may be affected by both the dose level and by the strain. It may be of further interest whether the sex of the mouse makes any difference in its response. Figure 7-1 shows how this data might be organized.

	Male			**Female**		
	Dose 1	Dose 2	Dose 3	Dose 1	Dose 2	Dose 3
Strain A	X_{111}	X_{112}	X_{113}	X_{211}	X_{212}	X_{213}
Strain B	X_{121}	X_{122}	X_{123}	X_{221}	X_{222}	X_{223}
Strain C	X_{131}	X_{132}	X_{133}	X_{231}	X_{232}	X_{233}
Strain D	X_{141}	X_{142}	X_{143}	X_{241}	X_{242}	X_{243}

Figure 7-1

In this design, there are three factors: dose, sex and strain. Each factor comprises several distinct groups called levels. In this case there are three levels of dose, two levels of sex and four levels of strain.

The questions of interest here are:

- Are there differences in response among dose levels regardless of sex and strain?
- Do male mice respond differently from female mice?
- Does it make a difference which strain the mice are from?

These questions refer to what are called main effects of the factors.

In addition, we may have questions related to the levels. For example, it may be that male mice respond differently than female mice, but how they respond may depend on the dose level. That is, the males' responses may be less than females' at lower dose levels and greater than females' at higher dose levels. This is known as an interaction of sex and dose. There are four possible factor interactions which may be investigated:

- a sex by dose interaction;
- a sex by strain interaction;
- a strain by dose interaction;
- and finally, a three-way dose by sex by strain interaction.

It is important to remember that only one kind of observation is being analyzed: blood pressure. The purpose of analysis of variance in such an experiment is to determine whether one or more of the different circumstances is having an important influence on the response.

Obviously, this experimental design is more complex than those illustrated in previous sections and will usually require more subjects (observations). But if it can be shown in the analysis that there is no significant difference among strains, then in future studies there will be no need to distinguish strains and they may be pooled into a single group. This can be extremely helpful when one strain is less readily obtained then others.

Analysis of variance isolates the effects of the different circumstances by subdividing the total sum of squares of all the observations about the grand mean into several components. The components correspond to each of the factors, plus all possible combinations of these factors (the interactions) and a component designated random or residual error. The estimated variation attributed to each factor or interaction is compared with that due to residual error. A particular factor or interaction is then labeled "statistically significant" when the variation is greater than the random error at some prescribed probability level.

7.2 PROPHET Usage: ANVAR

The public procedure ANVAR computes the analysis of variance for situations ranging from two to six factors. ANVAR requires data in an n-dimensional array (where n=2 to 6) and assumes that an n-factorial design with no replication is being used. (See Section 7.3 for instructions on making an array.) ANVAR computes the total sum of squares deviations from the overall mean, and divides this sum into all possible main effects and interactions.

These sums of squares are denoted as A, B, AB, C, AC, BC, etc., where each letter A, B, C, ect. is associated with a single factor. The main effect of the first factor is denoted by A (the variation in response for different levels of A, averaged over all levels of the remaining factors); AB is the interaction term between the first and second factor, ABC is the 3-factor interaction among the first three factors, etc.

A. Comparing ANVAR to Analysis of Variance in COMPARE

While ANVAR's approach has more subdivisions than most experimental designs require, it has the advantage that it can be adapted to a wide variety of designs. Compare, for example the output from ANVAR, Figure 7-2 with the output from COMPARE, Figure 4-6. The two-way design in this example is a single factor randomized blocks experiment.

COMPARE produces an analysis of variance table, where the source of variation is divided into GROUPS, BLOCKS and RESIDUALS. ANVAR, by contrast, returns an analysis of variance table with effects A, B, and AB. These are identical to the GROUPS, BLOCKS, and RESIDUALS, respectively. AB is the RESIDUALS term because the data were assigned randomly to each block; hence it is assumed that there is no block-by-treatment interaction. In other words, AB is assumed to contain variation due to the individual observations only, which is the random error in the experiment. (If you believe there is a block-by-treatment interaction, then replications should be included in the design so that estimates of a pure interaction term and pure error term can be made.)

If we had specified in COMPARE that the design was one-way, then the analysis of variance table would contain sums of squares for GROUPS and RESIDUALS terms only. ANVAR, however, always returns a table, with A, B, and AB, as in Figure 7-2, regardless of the design used. The GROUPS SS is the same as A. The RESIDUALS SS is computed by adding the sums of squares for B and AB. Similarly, the degrees of freedom for the RESIDUALS is the sum of the degrees of freedom for B and AB. The mean square is then found by dividing these two sums, RESIDUALS SS by RESIDUALS DF. Figure 7-3 shows the resulting table.

DIETANOV 3R X 3C

0 ROWNAME	1 SUM OF SQUARES	2 DEGREES OF FREEDOM	3 MEAN SQUARE
1. A	.393	4.	.09825
2. B	8.1535	3.	2.717833
3. AB	.779	12.	.06491667

Figure 7-2

Since the second factor is the replication in the one-way design, all the terms including B are pooled together to form the error term. Another way to look at this is to remember that the error sum of squares is equal to the total sum of squares minus the sum of squares due to treatments. Since the total sum of squares comprises terms A + B + AB, the error term must be B + AB.

DIETANOV2 2R X 3C

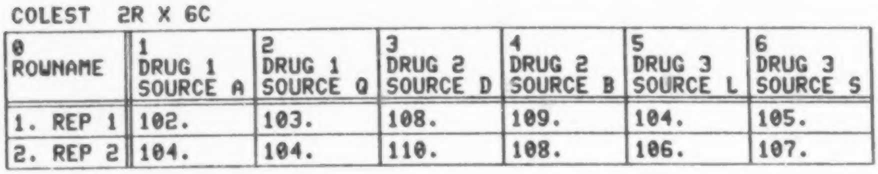
0 ROWNAME	1 SUM OF SQUARES	2 DEGREES OF FREEDOM	3 MEAN SQUARE
1. A	.393	4.	.09825
2. B + AB	8.9325	15.	.5955

Figure 7-3

Similar approaches can be taken with more complex designs. Table 7-2 at the end of Section 7 shows several common designs and how to pool terms together to form the final analysis of variance table. Table 7-3 shows the correspondence between the letters A, B, C and the formulae that compute them for a three dimensional design. The first column is the effect; the second, the corresponding sum of squares. The bottom portion shows examples of pooled effects and the resulting sum of squares after some algebraic manipulation. To illustrate the use of ANVAR and the information in Tables 7-2 and 7-3, we will use a one-way nested (hierarchical) design.

B. Nested design using ANVAR

The use of ANVAR is illustrated in an analysis of cholesterol level data, shown in Figure 7-4 (Zar, p. 144). Blood cholesterol concentrations are measured in women receiving three different drugs, where each drug has been obtained from two different sources. The response, cholesterol concentration, varies not only from the different drug group but also from the different sources. Let X_{ijk} denote the cholesterol concentration of the kth female, given the ith drug from the jth source.



COLEST 2R X 6C

0 ROWNAME	1 DRUG 1 SOURCE A	2 DRUG 1 SOURCE Q	3 DRUG 2 SOURCE D	4 DRUG 2 SOURCE B	5 DRUG 3 SOURCE L	6 DRUG 3 SOURCE S
1. REP 1	102.	103.	108.	109.	104.	105.
2. REP 2	104.	104.	110.	108.	106.	107.

Figure 7-4

Each response can be written as

$$X_{ijk} = \mu + \alpha_i + \beta_{(i)j} + \epsilon_{(ij)k} \qquad i = 1, 2, 3; j = 1, 2, 3, 4; k = 1, 2$$

That is, the concentration is divided into

- the overall mean μ ,
- the variation due to a drug α_i,
- the effect of source j for drug i $\beta_{(i)j}$.
- the differences of the individual persons $\epsilon_{(ij)k}$.

The assumptions that are made in this kind of analysis are that:

- The variation among sources is the same regardless of the kind of drug that is purchased, i.e., all 6 α_i have the same variance.
- The variation among women is the same no matter which drug is used or its source, i.e., all the 12 $\beta_{(i)j}$ have the same variance.

- The knowledge about β's gives us no information about the ε's. That is, knowing that for source A and drug 1 the responses in cholesterol are high tells us nothing about whether the first women's concentration is relatively high or low for that drug and source, i.e., β's and ε's are independent of one another.

- The β's are independent and normally distributed and the ε's are also independent and normally distributed.

ANVAR requires data stored in an n-dimensional FLOAT array. In this example n=3. Let COLEST be a FLOAT array array array such that COLEST (I,J,L) contains X_{ijk} described above. (A way of creating this array will be described later in Section 7-3.) The analysis of variance table COLANVAR in Figure 7-5, is the result of the commands

```
RAC PUBLIC ANVAR
CALL ANVAR(COLEST,3, 'COLANVAR')
```

COLANVAR 7R X 3C

0 ROWNAME	1 SUM OF SQUARES	2 DEGREES OF FREEDOM	3 MEAN SQUARE
1. A	61.16667	2.	30.58333
2. B	.3333333	1.	.3333333
3. AB	1.166667	2.	.5833333
4. C	5.333333	1.	5.333333
5. AC	1.166667	2.	.5833333
6. BC	1.333333	1.	1.333333
7. ABC	1.166667	2.	.5833333

Figure 7-5

Under nested one-factor designs (Table 7-2 at the end of this section) you will find how to pool the different terms to obtain the final analysis of variance table. Table 7-1 below summarizes the pertinent information for nested designs from Tables 7-2 and 7-3. The first column describes the source of variation. The second column contains the terms that you must pool in the analysis of variance table produced by ANVAR to obtain the sum of squares appropriate for the nested design. The third column in Table 7-1 shows the corresponding sum of squares that the terms in column 2 reduce to by algebraic manipulation.

TABLE 7-1. SUMS OF SQUARES FOR A ONE FACTOR, NESTED DESIGN.

Source of Variation	Pooled Terms	Sum of Squares
Groups (i.e., between groups)	A	$qr \sum_{i}^{p} (\bar{X}_{i..} - \bar{X}_{...})^2$
Subgroups (i.e., between subgroups within groups)	B + AB	$r \sum_{i}^{p} \sum_{j}^{q} (\bar{X}_{ij.} - \bar{X}_{i..})^2$
Error (i.e., within subgroups)	C + AC + BC + ABC	$\sum_{i}^{p} \sum_{j}^{q} \sum_{k}^{r} (X_{ijk} - \bar{X}_{ij.})^2$

Figure 7-6 shows the final analysis of variance table for this design, found by pooling sums of squares and degrees of freedom terms in COLANVAR as described in Table 7-1. The MEAN SQUARE column was found by dividing the SUM OF SQUARES by the DEGREES OF FREEDOM. The F-VALUE was found by dividing the MEAN SQUARE term by the appropriate ERROR term as described in Table 7-2. To find the significance level the system procedure %FSIGLEV, described in Part III, can be used.

ANOVA 3R X 5C

0 ROWNAME	1 SUM OF SQUARES	2 DEGREES OF FREEDOM	3 (D) MEAN SQUARE	4 F VALUE	5 SIGN. LEV.
1. DRUGS	61.16667	2.	30.58333	61.16666	.0037
2. SOURCES	1.5	3.	.5000001	.3333334	.802
3. ERROR	9.	6.	1.5		

Figure 7-6

In this example, the error MS measures the variability in cholesterol concentration from woman to woman and nothing else. The between source within drug subgroup MS measures the variability from the differences both between women and between sources. Finally, the group MS due to drugs depends on variation among women, variation among sources, and variation in response to the type of drug.

The first question we would want to test is "Do all the drugs have the same effect on cholesterol concentration?" Assume all the drugs had the same effect (the null hypothesis), then the group MS and subgroup MS would measure the same variation and hence should be about the same magnitude. By comparing the two, we can see whether there is any evidence that the two do not measure the same variation. We do this by comparing the group MS with the subgroup MS.

$$F = \frac{\text{Group MS}}{\text{Subgroup MS}} = \frac{30.58}{0.50} = 61.16$$

The probability of obtaining such a large F is less than 0.001 and hence the null hypothesis is rejected. There is very strong evidence that the three drugs effect cholesterol concentration differently.

The second hypothesis we can test is whether there is any differences among drug sources. This time we will compare the error MS with subgroup MS.

$$F = \frac{\text{Subgroup MS}}{\text{Error MS}} = \frac{0.50}{1.50} = 0.33$$

Since F has a significance level of .8, there appear to be no differences among sources. Thus, for a particular drug all the sampling variation seen in the data is from woman to woman.

C. Empties in ANVAR

ANVAR will allow up to 10% of the data, in an initially balanced experimental design, to be missing. Two different methods are used to estimate the missing values. First, if for each combination of factors there are replications, the missing values are replaced with the average of the observations in that combination. Otherwise a factorial design with no replication is assumed and the all factor interaction sum of squares is minimized. In both cases the degrees of freedom of the all factor interaction is reduced by one for each replacement.

By minimizing the error sum of squares the other sums of squares will be slightly overestimated. But unless the resulting F-values are near the border of significance, this overestimate is commonly neglected.

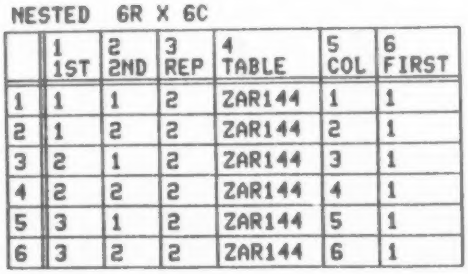
7.3 Preparing the Data for ANVAR using ARAFIL

As we noted earlier, ANVAR requires data in an n-dimensional array rather than in tables. When the data are in a two-way layout, moving the data to a two-dimensional array is easy.

But setting up an array with more than two dimensions can be tedious if you set one element at a time. The public procedure ARAFIL will aid you in setting up the array in two stages. The second stage will take the data from a table or several tables, known as the data table(s), and move the data into an array. The way in which the data are transferred is defined by a design table. This design table indicates where all the data are to be found and where in the data array they should go. The data can be arranged in data table(s) in any convenient way, except that the replicates must all be in either rows or columns. The first stage of ARAFIL helps in building the design table.

A. The design-table

Figure 7-7 shows the design table, NESTED, that was used to transfer the data in Figure 7-4 to the three-dimensional array, A, used by ANVAR.



NESTED 6R X 6C

	1 1ST	2 2ND	3 REP	4 TABLE	5 COL	6 FIRST
1	1	1	2	ZAR144	1	1
2	1	2	2	ZAR144	2	1
3	2	1	2	ZAR144	3	1
4	2	2	2	ZAR144	4	1
5	3	1	2	ZAR144	5	1
6	3	2	2	ZAR144	6	1

Figure 7-7

The first 3 columns of NESTED show the design characteristics of the data; the last 3 columns show where the data are to be found. Column 1 refers to the first factor, the different drugs, known by ANVAR as factor A. The second column refers to the source of the drugs, known by ANVAR as factor B. The sources have been coded as 1 and 2, such that sources A, D and L are assigned 1 and source Q, B, and S are assigned 2. The assignment is arbitrary except that for each drug there must be a source 1 and a source 2 (e.g., assigning A, B, and L to 1 and Q, D, and S to 2 would also be valid). The third column states how many replicates there are for each combination of drugs and sources. In this example two women are given the drug from a particular source. Since ANVAR can only be used when there is a balanced design (after filling for empties) the number of replicates must be the same for each combination of the factors.

The final three columns of NESTED describe where the data are found for the factor combinations in columns 1 and 2. The fourth column specifies the data table or data tables. The fifth column specifies which rows or columns of the data table the replications are in. The column heading "COL" shows that the replications are in columns. The last column in NESTED specifies the first replicate for the factor combination. For example, in row 3 of NESTED, the data being specified are the observations for drug 2 and the first source. Because there are two replicates, the data must be in rows 1 and 2 of column 3 of COLEST. (Empties could be in a row or column of replications.) Whether the replicates are in rows or columns is specified by the column heading of column 5. For this reason all replicates must be in columns only or rows only.

7.4 PROPHET Usage: ARAFIL

In the first stage, ARAFIL will create the design table with all but the last two or three columns filled. You will then leave ARAFIL and enter the information about where the data will be found. With the completed design table, you call ARAFIL again to fill the array with data.

The first question ARAFIL asks is whether you have a design table. If you do not, it will create one; otherwise, it will fill an array using the design table you name.

In the first stage you are prompted for the following information:

- The name the design table should have. This name must not have been used previously for another table.
- The dimension of the final data array. This is the number of factors, including any replication factors, in the design.
- The number of levels of each factor.
- The number of replicates (or blocks in a randomized block design).

With this information ARAFIL will then produce the design table filling in the first n columns, where n is the number of dimensions. The first n-1 columns will contain all combinations of the factors and the nth column will contain the number of replicates. Since ANVAR will only consider balanced designs the number of replicates will always be the same for each treatment combination.

The next two questions will be used for setting up the last three columns of the design table.

- Whether the replicates are in rows or columns.
- The name of the data table if there is only one.

The first question is used to name the second to last column "ROW" or "COL". If you have only one data table then the column "TABLE" will be filled with the table name you supply.

ARAFIL will stop at that point to allow you to fill in remaining portions of the table, which specify where the data can be found. You can use any method you choose to fill the table, such as ENTER ROWS, ENTER COLS, FILL etc.

Once you feel the design table is correctly made, call ARAFIL again to fill the data array using the design table. This second time you will be prompted for the following information:

- The name of the design table
- The name you want for the data array

The data array will be available for the remainder of the session.

Figure 7-8 shows a different layout for the colesterol data in Figure 7-4. The dialog used to obtain a design table for this new layout is shown in Figure 7-9. The resulting design table is in Figure 7-10; filling the array A2 is shown in Figure 7-11. The array A found using the above design is identical to the array A2 found using the design below.

COLEST2 4R X 3C

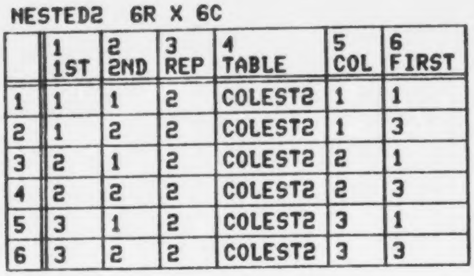
0 ROWNAME	1 DRUG 1	2 DRUG 2	3 DRUG 3
1. SOURCE 1	102.	108.	104.
2.	104.	110.	106.
3. SOURCE 2	103.	109.	105.
4.	104.	108.	107.

Figure 7-8

```
>RAC PUBLIC ARAFIL<
>CALL ARAFIL<

DO YOU HAVE A DESIGN TABLE WHICH SPECIFIES HOW TO FILL THE ARRAY?N<
NAME FOR DESIGN TABLE ?NESTED2<

HOW MANY DIMENSIONS WILL YOUR ARRAY CONTAIN ?3<
NUMBER OF LEVELS FOR DIMENSION 1?3<
NUMBER OF LEVELS FOR DIMENSION 2?2<
HOW MANY REPLICATIONS ARE THERE ?2<

ARE THE REPLICATES IN ROWS(R) OR COLS(C)?C<
ARE ALL THE DATA FROM THE SAME TABLE?Y<
WHAT IS THE NAME OF THE DATA TABLE?COLEST2<

YOU WILL NOW NEED TO FILL IN THE LAST 2 COLUMNS OF NESTED2
CALL ARAFIL AGAIN WHEN READY TO MAKE THE ARRAY
```

Figure 7-9

best copy available



NESTED2 6R X 6C

	1 1ST	2 2ND	3 REP	4 TABLE	5 COL	6 FIRST
1	1	1	2	COLEST2	1	1
2	1	2	2	COLEST2	1	3
3	2	1	2	COLEST2	2	1
4	2	2	2	COLEST2	2	3
5	3	1	2	COLEST2	3	1
6	3	2	2	COLEST2	3	3

Figure 7-10

```
>CALL ARAFIL<

DO YOU HAVE A DESIGN TABLE WHICH SPECIFIES HOW TO FILL THE ARRAY?Y<
WHAT IS THE NAME OF THE DESIGN TABLE?NESTED2<

HOW MANY DIMENSIONS WILL YOUR ARRAY CONTAIN?3<
WHAT NAME DO YOU WANT FOR THE DATA ARRAY?A2<
```

Figure 7-11

TABLE 7-2. REDUCTION OF ANALYSIS OF VARIANCE OF VARIOUS DESIGNS.

Description of Design	1 – Factor Nested	2 – Factor with Replication	2 – Factor Randomized Block	3 – Factor Randomized Blocks	Split Plot:	Split-Split Plot
Factor No. 1 2 3 4	Groups = A Subgroups = B Reps. = C	Rows = A Columns = B Reps. = C	Factor 1 = A Factor 2 = B Blocks = C	Factors A B C Blocks D	Main T. = A Sub T. = B Blocks = C	Main T. = A Sub T. = B Sub-sub T. = D Blocks = C
Combination for error term	Error = Sum of all terms with C	Error = C + AC + BC + ABC	Error = AC + ABC	Error = Sum of all terms with D	Error (a) = AC Error (b) = BC + ABC	Error (b) = BC + ABC Error (c) = CD + BCD + ACD + ABCD
Analysis of variance	Groups A Subgroups B + AB Error	Rows A Columns B Inter. AB Error	Factor 1 A Factor 2 B Inter. AB Blocks C Error	Main effects A B C 1st-order inter. AB AC BC 2nd order inter. ABC Error	Main T. A Blocks C Error (a) Sub T. B Inter. AB Error (b)	Main T. A Blocks C Error (a) AC Sub T. A Main × sub AB Error (b) Sub-sub T. D Main × sub-sub AD Sub × sub-sub BD 2nd-order inter. ABD Error (c)

TABLE 7-3. SUMS OF SQUARES FOR A 3-DIMENSIONAL DESIGN.

Source	Sum of Squares
A	$qr \sum_{i}^{p} (\overline{X}_{i..} - \overline{X}_{...})^2$
B	$pr \sum_{j}^{q} (\overline{X}_{.j.} - \overline{X}_{...})^2$
AB	$r \sum_{i}^{p} \sum_{j}^{q} (\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...})^2$
C	$pq \sum_{k}^{r} (\overline{X}_{..k} - \overline{X}_{...})^2$
AC	$q \sum_{i}^{p} \sum_{k}^{r} (\overline{X}_{i.k} - \overline{X}_{i..} - \overline{X}_{..k} + \overline{X}_{...})^2$
BC	$p \sum_{j}^{q} \sum_{k}^{r} (\overline{X}_{.jk} - \overline{X}_{.j.} - \overline{X}_{..k} + \overline{X}_{...})^2$
ABC	$\sum_{i}^{p} \sum_{j}^{q} \sum_{k}^{r} (X_{ijk} - \overline{X}_{ij.} - \overline{X}_{i.k} - \overline{X}_{.jk} + \overline{X}_{i..} + \overline{X}_{.j.} + \overline{X}_{..k} - \overline{X}_{...})^2$
total	$\sum_{i}^{p} \sum_{i}^{q} \sum_{k}^{r} (X_{ijk} - \overline{X}_{...})^2$
B + AB	$r \sum_{i}^{p} \sum_{j}^{q} (\overline{X}_{ij.} - \overline{X}_{i..})^2$
AC + BC + ABC	$\sum_{i}^{p} \sum_{i}^{q} \sum_{k}^{r} (X_{ijk} - \overline{X}_{ij.} - \overline{X}_{..k} + \overline{X}_{...})$
C + AC + BC + ABC	$\sum_{i}^{p} \sum_{j}^{q} \sum_{k}^{r} (X_{ijk} - \overline{X}_{ij.})^2$

8. Covariance

8. Covariance

SECTION 8

ANALYSIS OF COVARIANCE

8.1 Introduction

In Section 4 we discussed the technique of analysis of variance for testing differences among sample means. Section 6 described the methods of regression analysis to study the relation between a dependent variable and an independent variable. Covariance analysis is a combination of these procedures. It is designed to investigate and compare several regressions simultaneously. The X's are known as the covariates. This section will discuss covariance analysis only as it applies to simple linear regression.

The fundamental purpose of analysis of covariance is to determine whether several distinct sample regression lines could all have been drawn from a common population. If this null hypothesis is rejected, it is of further interest to determine which pairs of regression lines are different, and in what way, either by slope or by intercept. In addition, covariance analysis is often useful for removing the effect of one or more unwanted factors.

For example, assume the growth of three plants, each given a different fertilizer, is plotted versus time. In this instance the slopes of the regression lines represent growth rate, and it is desired to determine if there are any differences among the growth rates attributable to the three fertilizers. This could be studied by comparing the slopes using the analysis of covariance. In this situation the slopes of the regression lines are of primary interest.

In other situations, the intercepts are of major interest. For example, a study might be designed to compare the development of reading skills in three classes of children. Since it would be desirable to remove the influence of each child's IQ, the final sets of observations (say, scores on a reading test) could be plotted against IQ, and three regression lines found. These lines would then be compared using the analysis of covariance.

8.2 Basic Models in Analysis of Covariance

The underlying assumptions in the analysis of covariance are essentially those that apply to regression analysis in general. For every value of X, there is assumed a population of Y values which are normally distributed with equal variances and have means which lie on the true regression line. This model may be expressed as

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij}$$

where, as in regression generally, the error terms ϵ_{ij} are independent and normally distributed with means equal to zero and common variance. Here j represents the observation, and i the group (or line).

Consider the case of three regression lines, shown in Figure 8-1.

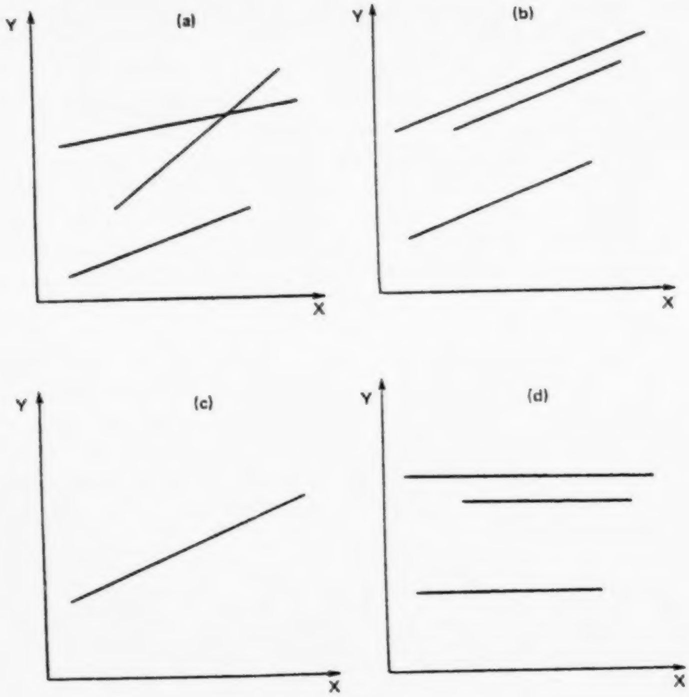


Figure 8-1

Figure 8-1 depicts four possible situations. In (a), there are three distinct regression lines with unique slopes and intercepts. In this case, it is difficult to make comparisons among the Y-values for the three lines without specifying the corresponding X-values. The situation in (b) is different, in that the lines are parallel although they differ in elevation (i.e., their intercepts are different). This may be expressed in the model as

$$Y_{ij} = \alpha_i + \beta X_{ij} + \epsilon_{ij}$$

where now there is a single slope, β , common to all three lines. In this case the differences in the Y-population means for each of the three lines is simply the difference in their intercepts. In (c) there is no difference among the three lines, and in fact, all the data can best be described by a single regression equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \epsilon_{ij}$$

with single values for α and β .

Finally, in (d) the three lines are parallel with zero slope. In this case the X's have no effect on the Y's; hence, an analysis of variance of the Y's would be sufficient.

$$Y_{ij} = \alpha_i + \epsilon_{ij}$$

Three hypotheses can be tested by comparing fits of situation (a) with (b), (b) with (c), and (b) with (d). If the fit of the data becomes significantly worse by going from one situation to another, the model with more parameters must be retained as the most accurate.

8.3 Computations in Analysis of Covariance

Consider our example of 3 regression lines under the null hypothesis that all three lines are best represented by three distinct regression lines: that is, by three pairs of slopes and intercepts. In this case we may compute a residual SS term (see Section 6) for each line as

$$\text{Residual SS}_i = \sum_j (Y_{ij} - (a_i + b_i X_{ij}))^2$$

When these three residual SS terms are summed they form the "pooled residual SS". Under the second null hypothesis, the slopes of the three lines are not different, but rather each line can best be represented by a common estimate of the slope (b_C). Thus, we have a "Common residual SS", given by

$$\text{Common SS} = \sum\sum (Y_{ij} - (a'_i + b_c X_{ij}))^2$$

$$b_c = \frac{\sum\sum (X_{ij} - \overline{X}_{i\cdot})(Y_{ij} - \overline{Y}_{i\cdot})}{\sum\sum (X_{ij} - \overline{X}_{i\cdot})^2}$$

If the three data sets can best be represented by a single line, then the regression equation is formed by considering all the data to have come from a single sample. In this case we form the "Total Residual SS" as

$$\text{Total SS} = \sum\sum (Y_{ij} - (a_T + b_T X_{ij}))^2$$

where b_T is the slope computed in the usual way on the entire set of data.

In order to compare the means from the different groups an adjusted mean Y may be calculated at the overall mean $\overline{X}_{\cdot\cdot}$, that is,

$$\text{adj } Y_i = \overline{Y}_i + b_c(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})$$

8.4 Hypothesis Testing in Analysis of Covariance

The various null hypotheses in the analysis of covariance are tested in a manner analagous to the overall test of equality of means in the analysis of variance (see Section 4). That is, an F statistic is computed as the ratio of a specific mean square term and the appropriate error mean square.

The first hypothesis to be tested is that of equality of slopes: that is, to determine if $\beta_1=\beta_2..=\beta_n$ can be accepted as true, as opposed to the alternative hypothesis that not all slopes are the same. Rejection of the null hypothesis does not specify which slopes are different; it merely indicates that some differences do exist.

To test the null hypothesis two models, one allowing the slopes to vary and the other forcing the slopes to be equal, are compared. In this case we may compute a mean square term due to the assumption of equality of k slopes by considering the difference in residual sum of square terms for the two models:

Mean Square (Equality of Slopes) = (Common SS-Pooled SS)/(k-1)

and a residual or error mean square as

Error Mean Square = Pooled SS/(N-2k)

where N is the total number of (X,Y) pairs. Then

$$F = \frac{(\text{Common SS} - \text{Pooled SS})/(k-1)}{\text{Pooled SS}/(\sum n_i - 2k)}$$

When the slopes are in fact different we would expect the residual variability to be much larger when a common slope is assumed than when the lines are considered to have separate slopes. This is equivalent to saying that the common SS is much greater than the pooled SS, relative to the random error (pooled MS). Thus the above F-ratio will be large, indicating rejection of the null hypothesis of equality of slopes.

If this calculated F does not exceed the critical value of the F-distribution, then the null hypothesis would be accepted. In this case, all lines share a common slope, which is given above.

If the null hypothesis is rejected, you may want to perform a multiple range comparison procedure to determine specifically which pairs of slopes are different. This is analogous to performing either the Newman-Keuls or Dunnett's test following the analysis of variance to identify differences among means. If all pairwise comparisons are being made, the following statistic is employed

$$q = \frac{b_A - b_B}{SE}$$

$$SE = \sqrt{\frac{(s^2_{y.x})_p}{2}\left(\frac{1}{\Sigma\,(X_{A_i} - \overline{X}_{A\cdot})^2} + \frac{1}{\Sigma\,(X_{B_i} - \overline{X}_{B\cdot})^2}\right)}$$

$$(s^2_{y.x})_p = \frac{\text{Pooled SS}}{\Sigma\, n_i - 2k}$$

where A and B indicate any two regression lines. The calculated value of q is compared to the critical value of the Studentized Range using the Newman-Keuls method.

If one of the lines is considered to be the control, then the test statistic is the same as that for all pairwise comparisons except that

$$\text{SE for control} = \sqrt{2}\ \ (\text{SE for all pairwise})$$

and the test statistic is compared to Dunnett's critical values.

If the null hypothesis of equality of slopes is accepted, you then make an overall test for the equality of elevations. This null hypothesis states that there are no differences among the elevations, or intercepts, given that the lines share a common slope, i.e., $\alpha_1 = \alpha_2 \ldots = \alpha_n$.

In this case we form a mean square term due to equality of intercepts as

Mean Square (intercepts) = (Total SS - Common SS)/(k-1)

with the error mean square given by

Error Mean Square = Common SS/(N-k-1)

When the intercepts are in fact different, you would expect to see a large difference between the total SS (where the model assumes only one line fit to the data) and the common SS, and the following F-ratio will be large

$$F = \frac{(\text{Total SS} - \text{Common SS})/(k - 1)}{\text{Common SS}/(\sum n_i - k - 1)}$$

When F exceeds the critical value of the F-distribution, the null hypothesis is rejected, in which case a multiple range comparison procedure may be performed to determine which pairs of intercepts are different. For the case of all pairwise comparisons, the test statistic is

$$q = \frac{(\overline{Y}_{A\cdot} - \overline{Y}_{B\cdot}) - b_c(\overline{X}_{A\cdot} - \overline{X}_{B\cdot})}{SE}$$

$$SE = \sqrt{\frac{(s^2_{y.x})_c}{2}\left(\frac{1}{n_A} + \frac{1}{n_B} + \frac{(\overline{X}_{A\cdot} - \overline{X}_{B\cdot})^2}{\sum(X_{Aj} - \overline{X}_{A\cdot})^2 + \sum(X_{Bj} - \overline{X}_{B\cdot})^2}\right)}$$

$$(s^2_{y.x})_c = \frac{\text{Common SS}}{\sum n_i - k - 1}$$

where A and B denote any two lines and q is compared to critical values of the Studentized Range.

If one line is considered the control, the SE is $\sqrt{2}$ times the SE for all pairwise comparisons.

Finally, to test the null hypothesis that there is no dependence of Y on X, we form the F-ratio of

Mean Square (Y independent) = $\Sigma\Sigma(Y_{ij}-\overline{Y}_{..})^2$ - Common SS

with

Error Mean Square = Common SS/(N-k-1)

$$F = \frac{\sum\sum(Y_{ij} - \overline{Y}_{..})^2 - \text{Common SS}}{\text{Common SS}/(\sum n_i - k - 1)}$$

It should be clear that all the above hypothesis tests follow a similar line of reasoning. That is, two models are hypothesized and the residual sums of squares for each model is calculated. Next, a mean square term is formed by finding the differences in the two residual sums of squares and dividing the differences by the change in the degrees of freedom for the two residual sums of squares. The error mean square term is the residual mean square term for the model that restricts the parameters the least. The F-ratio is calculated by comparing the two mean square terms.

8.5 PROPHET Usage: TEST LINES

Covariance analysis is performed in PROPHET with the TEST LINES command. The different sets of X and Y data to be compared are contained in the columns of one or more PROPHET tables. When you use TEST LINES, you are first asked "DO THE DATA COME FROM MORE THAN 1 TABLE?". If you respond "YES" you are then asked for the number of data tables; otherwise, one table is assumed. Following this you enter the name of the table.

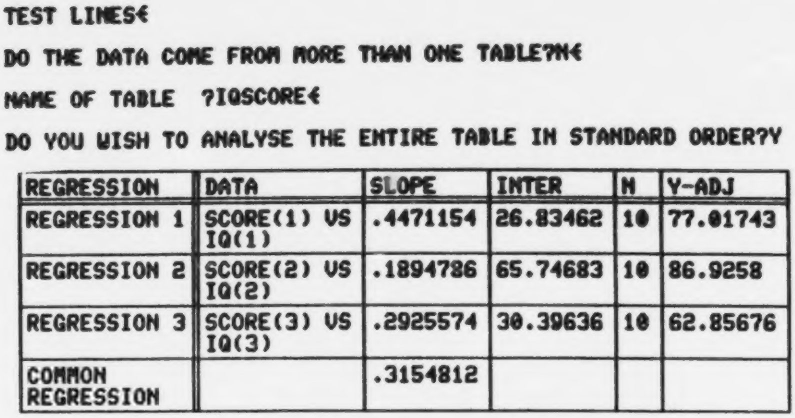
Standard Order

In many instances the data table will be organized such that columns 1 and 2 contain the X and Y data respectively for the first regression, columns 3 and 4 contain the X and Y data for the second regression, etc. This type of data organization is referred to in the TEST LINES dialogue as "standard order". Thus, for each table, you are asked, "DO YOU WISH TO ANALYZE THE ENTIRE TABLE IN STANDARD ORDER?". If you answer "YES" such an arrangement is assumed, and you need not specify explicitly which columns contain the data. If you respond "NO", you are asked to specify the columns for each pair of X and Y data. Note that in either case the columns must be of type "NUMBER". If any X or Y value is missing, that pair of data is ignored in the analysis. An example of this dialogue is shown in Figure 8-2.

```
TEST LINES<

DO THE DATA COME FROM MORE THAN ONE TABLE?N<

NAME OF TABLE  ?IQSCORE<

DO YOU WISH TO ANALYSE THE ENTIRE TABLE IN STANDARD ORDER?N<

ENTER NUMBER OF REGRESSION LINES FOR THIS TABLE?3<
REGRESSION 1
        COLUMN NUMBER FOR X-DATA?1<
        COLUMN NUMBER FOR Y-DATA?2<
REGRESSION 2
        COLUMN NUMBER FOR X-DATA?1<
        COLUMN NUMBER FOR Y-DATA?3<
REGRESSION 3
        COLUMN NUMBER FOR X-DATA?1<
        COLUMN NUMBER FOR Y-DATA?4<
```

Figure 8-2

After you have entered all the information pertaining to the source of the data, TEST LINES performs each regression and displays a table such as that in Figure 8-3. The regressions are numbered in sequence in the rows; the last row contains the estimate of the common slope. The last column contains the adjusted Y's for each line.



```
TEST LINES←

DO THE DATA COME FROM MORE THAN ONE TABLE?N←

NAME OF TABLE  ?IQSCORE←

DO YOU WISH TO ANALYSE THE ENTIRE TABLE IN STANDARD ORDER?Y
```

REGRESSION	DATA	SLOPE	INTER	N	Y-ADJ
REGRESSION 1	SCORE(1) VS IQ(1)	.4471154	26.83462	10	77.01743
REGRESSION 2	SCORE(2) VS IQ(2)	.1894786	65.74683	10	86.9258
REGRESSION 3	SCORE(3) VS IQ(3)	.2925574	30.39636	10	62.85676
COMMON REGRESSION		.3154812			

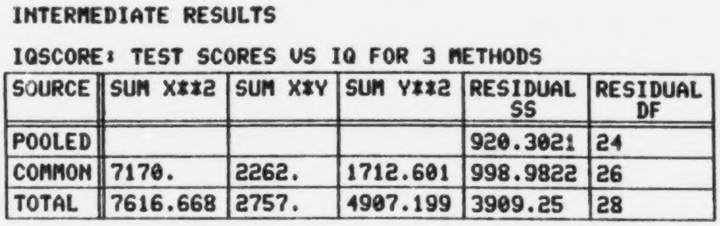
Figure 8-3

Control Group

Next, you are asked if there is a control group. If you respond "YES" you then enter the regression number from the regressions table (see Figure 8-3) for the control group.

Intermediate Results

If you type "YES" in response to the question "DO YOU WISH TO SEE THE INTERMEDIATE RESULTS?" a table such as that in Figure 8-4 will be displayed.



INTERMEDIATE RESULTS

IQSCORE: TEST SCORES VS IQ FOR 3 METHODS

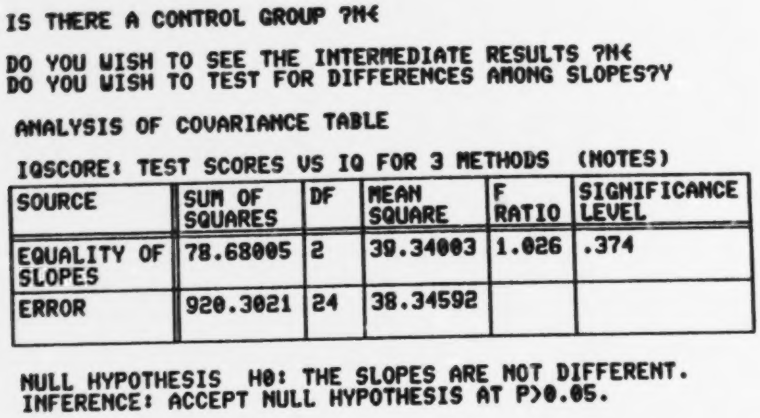
SOURCE	SUM X**2	SUM X*Y	SUM Y**2	RESIDUAL SS	RESIDUAL DF
POOLED				920.3021	24
COMMON	7170.	2262.	1712.601	998.9822	26
TOTAL	7616.668	2757.	4907.199	3909.25	28

Figure 8-4

(Note that in the "SUM" columns the X and Y values are actually computed as $(X-\bar{X})$ and $(Y-\bar{Y})$, i.e., the data are corrected for their means).

Slopes Comparison

You are then asked if you wish to test for differences among slopes. If you respond "YES", an analysis of covariance table is displayed, as in Figure 8-5.

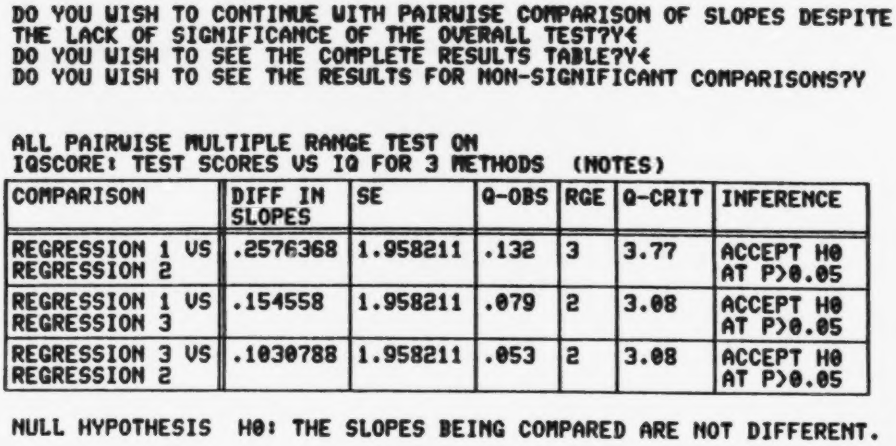


```
IS THERE A CONTROL GROUP ?N<

DO YOU WISH TO SEE THE INTERMEDIATE RESULTS ?N<
DO YOU WISH TO TEST FOR DIFFERENCES AMONG SLOPES?Y

ANALYSIS OF COVARIANCE TABLE

IQSCORE: TEST SCORES VS IQ FOR 3 METHODS  (NOTES)
+-------------+-----------+---+-----------+-------+-------------+
|SOURCE       |SUM OF     |DF |MEAN       |F      |SIGNIFICANCE |
|             |SQUARES    |   |SQUARE     |RATIO  |LEVEL        |
+-------------+-----------+---+-----------+-------+-------------+
|EQUALITY OF  |78.68005   |2  |39.34003   |1.026  |.374         |
|SLOPES       |           |   |           |       |             |
+-------------+-----------+---+-----------+-------+-------------+
|ERROR        |920.3021   |24 |38.34592   |       |             |
+-------------+-----------+---+-----------+-------+-------------+

NULL HYPOTHESIS  H0: THE SLOPES ARE NOT DIFFERENT.
INFERENCE: ACCEPT NULL HYPOTHESIS AT P>0.05.
```

Figure 8-5

The statement of the null hypothesis, pertinent test statistics, and inference are presented in the usual fashion.

If the resulting F-statistic indicates that differences exist among the slopes, TEST LINES performs the multiple comparison procedure automatically. If the inference is that the null hypothesis is to be accepted, you are asked "DO YOU WISH TO CONTINUE WITH PAIRWISE COMPARISON OF SLOPES DESPITE THE LACK OF SIGNIFICANCE OF THE OVERALL TEST?". You must answer "YES" if you want multiple comparisons to be made. When the multiple comparisons are performed, you are asked if you wish to see the complete results table and the results for nonsignificant comparisons. The complete results table for our example is shown in Figure 8-6.



```
DO YOU WISH TO CONTINUE WITH PAIRWISE COMPARISON OF SLOPES DESPITE
THE LACK OF SIGNIFICANCE OF THE OVERALL TEST?Y<
DO YOU WISH TO SEE THE COMPLETE RESULTS TABLE?Y<
DO YOU WISH TO SEE THE RESULTS FOR NON-SIGNIFICANT COMPARISONS?Y

ALL PAIRWISE MULTIPLE RANGE TEST ON
IQSCORE: TEST SCORES VS IQ FOR 3 METHODS  (NOTES)
```

COMPARISON	DIFF IN SLOPES	SE	Q-OBS	RGE	Q-CRIT	INFERENCE
REGRESSION 1 VS REGRESSION 2	.2576368	1.958211	.132	3	3.77	ACCEPT H0 AT P>0.05
REGRESSION 1 VS REGRESSION 3	.154558	1.958211	.079	2	3.08	ACCEPT H0 AT P>0.05
REGRESSION 3 VS REGRESSION 2	.1030788	1.958211	.053	2	3.08	ACCEPT H0 AT P>0.05

```
NULL HYPOTHESIS  H0: THE SLOPES BEING COMPARED ARE NOT DIFFERENT.
```

Figure 8-6

If you do not ask for the complete results table, only the columns "COMPARISON", "DIFF IN SLOPES" and "INFERENCE" will be displayed.

Intercepts Comparison

If you do not perform the test for differences among slopes, or if the test indicates that the slopes are not different, you are asked if you wish to test for differences among intercepts. If you respond "YES", an analysis of covariance table such as that shown in Figure 8-7 will be displayed.

DO YOU WISH TO TEST FOR DIFFERENCES AMONG INTERCEPTS?Y<

ANALYSIS OF COVARIANCE TABLE

IQSCORE: TEST SCORES VS IQ FOR 3 METHODS (NOTES)

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	SIGNIFICANCE LEVEL
EQUALITY OF INTERCEPTS	2910.268	2	1455.134	37.872	.0001
ERROR	998.0822	26	38.42239		

NULL HYPOTHESIS H0: THE INTERCEPTS ARE NOT DIFFERENT.
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.

Figure 8-7

You can then perform tests for specific pairwise differences among intercepts in the usual way as shown in Figure 8-8.

ALL PAIRWISE MULTIPLE RANGE TEST ON
IQSCORE: TEST SCORES VS IQ FOR 3 METHODS (NOTES)

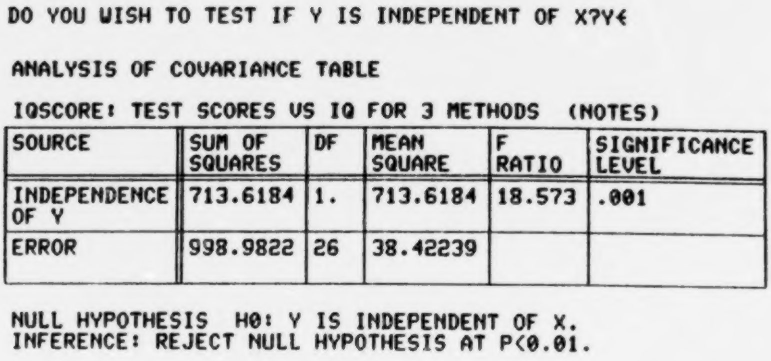
COMPARISON	DIFF IN INTERCEPTS	SE	Q-OBS	RGE	Q-CRIT	INFERENCE
REGRESSION 2 VS REGRESSION 3	24.06904	1.964352	12.253	3	4.727	REJECT H0 AT P<0.01
REGRESSION 2 VS REGRESSION 1	9.908368	2.011204	4.927	2	4.09	REJECT H0 AT P<0.01
REGRESSION 1 VS REGRESSION 3	14.16067	2.03767	6.949	2	4.09	REJECT H0 AT P<0.01

NULL HYPOTHESIS H0: THE INTERCEPTS BEING COMPARED ARE NOT DIFFERENT.

Figure 8-8

Test for Independence

Finally, after a comparison of intercepts, you are asked if you wish to test if the Y-data are independent of the X-data. If you respond "YES", the results shown in Figure 8-9 will be displayed.



```
DO YOU WISH TO TEST IF Y IS INDEPENDENT OF X?Y<

ANALYSIS OF COVARIANCE TABLE

IQSCORE: TEST SCORES VS IQ FOR 3 METHODS  (NOTES)
+--------------++----------+----+----------+--------+-------------+
|SOURCE        ||SUM OF    |DF  |MEAN      |F       |SIGNIFICANCE |
|              ||SQUARES   |    |SQUARE    |RATIO   |LEVEL        |
+==============++==========+====+==========+========+=============+
|INDEPENDENCE  ||713.6184  |1.  |713.6184  |18.573  |.001         |
|OF Y          ||          |    |          |        |             |
+--------------++----------+----+----------+--------+-------------+
|ERROR         ||998.9822  |26  |38.42239  |        |             |
|              ||          |    |          |        |             |
+--------------++----------+----+----------+--------+-------------+

NULL HYPOTHESIS  H0: Y IS INDEPENDENT OF X.
INFERENCE: REJECT NULL HYPOTHESIS AT P<0.01.
```

Figure 8-9

9. Miscellaneous

9. Miscellaneous

SECTION 9

MISCELLANEOUS TESTS

9.1 Multivariate Analysis

Sections 1 to 8 discuss statistical tests which are applied to a single variate or characteristic. In analysis of variance one response variable, which is divided into classes for each combination of factors, is analyzed. In regression and covariance analysis, there is only one random variable, the dependent variable. These tests are univariate tests.

In a multivariate statistical test several characteristics are analyzed simultaneously. For example, suppose you wanted to measure heart rate, blood pressure and cholesterol level of males and females (Figure 9-1). You might want to determine whether there are any differences between males and females. To do this, you could apply separate t-tests on each variable and from these tests decide whether males and females differ.

However, with such a design you must also consider whether these variables are correlated. The multivariate test Hotelling's T^2 will compare these three characteristics with one test, taking into consideration their variances and covariances. This is an extension of the two-sample Student's t-test to more than one variate. (This multivariate extension is not to be confused with the multisample generalization of the t-test in analysis of variance, in which a single variable under a number of circumstances is studied.)

FEMALES 6R X 3C

	1 HEART RATE	2 BLOOD PRESSURE	3 CHOLESTEROL LEVEL
1	75.	95.	102.
2	85.	95.	80.
3	68.	95.	110.
4	90.	93.	87.
5	83.	105.	107.
6	74.	98.	96.

MALES 8R X 3C

	1 HEART RATE	2 BLOOD PRESSURE	3 CHOLESTEROL LEVEL
1	66.	96.	90.
2	75.	110.	120.
3	75.	95.	101.
4	72.	95.	99.
5	81.	92.	99.
6	85.	105.	104.
7	69.	95.	105.
8	83.	103.	108.

Figure 9-1

9.2 Hotelling's T^2

Hotelling's T^2 tests whether the means of two multivariate normal populations are the same, assuming that the covariance matrices are equal. Let $X_{ij}^{(k)}$ denote the jth observation on the kth variate for population i. The null hypothesis is

$$H_0: \underline{\mu}_1 = \underline{\mu}_2$$

where

$$\underline{\mu}_i = (\mu_i^{(1)}, \mu_i^{(2)}, \mu_i^{(3)}, \ldots, \mu_i^{(p)}),\ i = 1,2$$

A vector of sample mean differences is used so that the kth elements of the vector $\underline{D}$ is

$$d^{(k)} = \overline{X}_1^{(k)} - \overline{X}_2^{(k)}$$

where

$$\overline{X}_1^{(k)} = \sum_{j=1}^{n_1} X_{1j}^{(k)}/n_1$$

and

$$\overline{X}_2^{(k)} = \sum_{j=1}^{n_2} X_{2j}^{(k)}/n_2$$

If the covariance matrices S_1 and S_2 can be considered the same they can then be pooled to form a common covariance matrix S by the following formula:

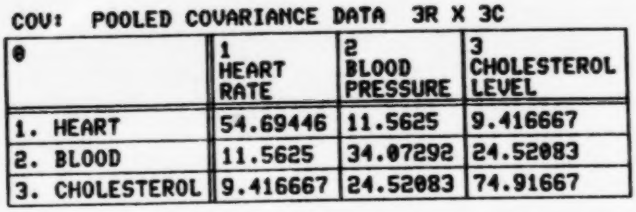
$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

That is, element S_{kk} and S_{km} in S are

$$S_{kk} = \frac{\Sigma(X_{ij}^{(k)} - \overline{X}_1^{(k)})^2 + \Sigma(X_{2j}^{(k)} - \overline{X}_2^{(k)})^2}{n_1 + n_2 - 2}$$

$$S_{km} = \frac{\Sigma(X_{ij}^{(k)} - \overline{X}_1^{(k)})(X_{1j}^{(m)} - \overline{X}_1^{(m)}) + \Sigma(X_{2j}^{(k)} - \overline{X}_2^{(k)})(X_{1j}^{(m)} - \overline{X}_2^{(m)})}{n_1 + n_2 - 2}$$

Figure 9-2 shows the pooled covariance matrix for the data in Figure 9-1.



COV: POOLED COVARIANCE DATA 3R X 3C

0	1 HEART RATE	2 BLOOD PRESSURE	3 CHOLESTEROL LEVEL
1. HEART	54.69446	11.5625	9.416667
2. BLOOD	11.5625	34.07292	24.52083
3. CHOLESTEROL	9.416667	24.52083	74.91667

Figure 9-2

The scalar statistic T^2 is computed from the vector of differences of means $\underline{D}$ and the pooled covariance matrix S as

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) \underline{D}\, S^{-1} \underline{D}$$

The statistic is distributed as the T^2 distribution with n_1+n_2-2 degrees of freedom. Under the null hypothesis

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\; T^2$$

is distributed as F with p and $n_1 + n_2 - p - 1$ degrees of freedom.

9.3 PROPHET Usage: HOTELLING

The public procedure HOTELLING will compute Hotelling's T^2 to test whether two multivariate normal samples are from the same population. The data for each sample must be in a separate table so that there is a column for each characteristic. The columns in the two tables must correspond so that the k<u>th</u> column in both tables contains data for the same characteristic. The number of rows in each table can be different but no row can contain empties. To call HOTELLING, type the following commands:

```
RAC PUBLIC HOTELLING
CALL HOTELLING
```

You will be asked for the names of the data tables. You will next be asked whether you want to use all the columns in the table. If you do not, you will be asked to type in the columns to be included. The T^2 statistic and its significance level will be reported. Figure 9-3 shows the analysis of the data in Figure 9-1.

```
>RAC PUBLIC HOTELLING< >
CALL HOTELLING<
NAME OF FIRST DATA TABLE?FEMALES<
NAME OF SECOND DATA TABLE?MALES<
DO YOU WANT TO INCLUDE ALL CHARACTERISTICS?Y<

HOTELLING'S STATISTIC: 3.984195
SIGNIFICANCE LEVEL: .391


HOTELLING COMPLETED.  TOOK 4.751001 SECONDS.
```

Figure 9-3

Appendices
and Index

Appendices
and Index

APPENDIX A:

FUNCTIONS AND PROCEDURES

Introduction

In addition to the statistical commands discussed in Part II, PROPHET makes a number of statistical tests available through calls to statistical functions and procedures. Most of these procedures provide the same facilities that are available within the commands, but in a form suitable for use from procedures which you write i.e., they do not ask questions or require answers. These procedures are briefly described in this part, along with a description of the arguments. In most cases the test statistic and its significance level are returned. The exact significance level is computed whenever possible, but often for small samples only tabulated values are available. In these instances the significance level is one of the following: 1, .05, .01, corresponding to a significance level between 1 and .05, .05 and .01 and less than .01, respectively.

1. %ANOVA

%ANOVA performs one- or two-way analysis of variance, followed by either the Newman-Keuls or Dunnett's multiple comparison procedure. (See Section 4.)

```
%ANOVA(data,row-col,control-index,all-stats,nonsigs,block-spec)
```

Argument	Type	I/O	Meaning/Value
data	TABLE	Input	Source for data.
row-col	TEXT	Input	"COLUMN" or "C" for columns; "ROW" or "R" for rows (if empty defaults to columns)
control-index	FIXED	Input	Row (column) index for control group (empty if no control group).
all-stats	BOOL	Input	TRUE to obtain complete results table; FALSE for compact results table.
nonsigs	BOOL	Input	TRUE to obtain results for nonsignificant comparisons; FASLE to obtain only significant results.
block-spec	TEXT	Input	"BLOCKED" or "B" for blocked data (2-way), empty otherwise (1-way).

Display: ANOVA table and multiple comparison results table.

For other multiple comparison tests see %DOKSUM, %FRIEDMAN and %KRUSKALWALLIS.

2. %CHISIGLEV

Returns the upper one-tail significance level of X, where X is chi-squared.

```
level = %CHISIGLEV (df,X)
```

Argument	Type	I/O	Meaning/Value
level	FLOAT	Output	Significance level
df	FIXED	Input	Degrees of freedom
X	FLOAT	Input	Chi-squared variate

The lower one-tail signficance level is 1.-%CHISIGLEV(df,X).

Display: none.

3. CONFCURVSQT

CONFCURVSQT adds confidence curves to a fitted line on a graph for some specified confidence level. It is the quiet entry to the public procedure CONFCURVS and therefore you must first RAC PUBLIC CONFCURVS before using it. (See Section 6.)

```
CONFCURVSQT(data, data-curve-num, fitted-curve-num, conf-lev)
```

Argument	Type	I/O	Meaning/Value
data	GRAPH	Input	Name of graph to which you want confidence curves added.
data-curve-num	FIXED	Input	The index of the curve of data points.
fitted-curve-num	FIXED	Input	The index of the curve which is the fitted line.
conf-lev	FLOAT	Input	Confidence level between 0.0 and 1.0 exclusive.

Display: None.

4. %DISPTEST

Tests the equality of dispersions of two samples (equivalent to the TEST DISPERSIONS command). (See Section 3.)

```
%DISPTEST(data1,data2,adjust-flg,stat,sig-lev)
```

Argument	Type	I/O	Meaning/Value
data1,data2	FLOAT array	Input	The two sets of data for which the dispersions are to be compared.
adjust-flg	BOOL	Input	Set to TRUE to have adjustment for median, FALSE for no adjustment.
stat	FLOAT	Output	Computed test-statistic.
sig-lev	FLOAT	Output	Significance level of "stat".

Display: none.

For parametric test see %FTEST.

5. %DISTRIBUTION

Tests an array of data against a known distribution using the Kolmorgov-Smirnov test. (See Section 2.)

```
test-result = %DISTRIBUTION(data,distrib-type,par1,par2,stat,sig-lev)
```

Argument	Type	I/O	Meaning/Value
test-result	TEXT	Output	Text of results formatted so it can be displayed.
data	FLOAT array	Input	Data to be tested.
distrib-type	TEXT	Input	Contains the name of the distribution that you wish to compare to. Only the first letter is required (see below).
par1,par2	FIXED or FLOAT	Input/Output	Parameters of the distribution. If empty, estimated from the data (see below).
stat	FLOAT	Output	Calculated test statistic.
sig-lev	FLOAT	Output	Significance level of the test statistic.

Display: none.

The usage of "par1" and "par2" for different distributions are as follows:

TYPE OF DISTRIBUTION	USAGE OF PAR1 AND PAR2
normal (N)	"par1" and "par2" are the values of the population mean and variance, respectively. The default option is EMPTY, in which case the sample mean and variance will be used.
bionomial (B)	"par1" is the total number of trials and "par2" is the probability of success on each trial (p), for the hypothetical population. The default option is EMPTY, in which case they will be estimated from functions of the sample mean and variance.
uniform (U)	"par1" and "par2" are the lower and upper bounds, respectively. Again, the default option is EMPTY, in which case the minimum and maximum values of the sample are used.
χ^2 (C)	"par1" is the degrees of freedom of the distribution. "Par2" is not used, and should be EMPTY. If "par1" is also EMPTY, then the degrees of freedom is estimated from the sample mean.
Poisson (P)	"par1" is the parameter m of the Poisson distribution. If EMPTY, it will be estimated from the sample mean. "par2" must be EMPTY.
t (T)	"par1" is the degrees of freedom of the distribution of the population. If set to EMPTY, the degrees of freedom will be estimated as a function of the sample variance. "Par2" should be set to EMPTY.
F	"par1" and "par2" are the degrees of freedom in the numerator and denominator, respectively. If EMPTY, they will be estimated from the sample mean and variance.

6. %DOKSUM

Performs Doksum's nonparametric test for blocked data. First tests the overall significance then performs a multiple comparison test for either all pairwise comparisons or all comparisons to a control.

```
%DOKSUM(data,row-col,control-index,all-stats,non-sigs)
```

Argument	Type	I/O	Meaning/Value
data	TABLE	Input	Source for data.
row-col	TEXT	Input	"COLUMN" or "C" for columns; "ROW" or "R" for rows (if empty defauls to columns)
control-index	FIXED	Input	Row (column) index for control group (empty if no control group).
all-stats	BOOL	Input	TRUE to obtain all statistics in RESULTS TABLE; FALSE results in compact results table.
nonsigs	BOOL	Input	TRUE to obtain results for nonsignificant comparisons; FALSE to obtain only significant results.

Display: Overall test and multiple comparison results table.

For other multiple comparison tests see %ANOVA, %FRIEDMAN and %KRUSKALWALLIS.

7. %FITFUNQT

%FITFUNQT performs a nonlinear regression on one independent variable using MLAB, (equivalent to FIT FUNCTION). (See Section 6.)

```
fit-text =   %FITFUNQT(data-tab, x-col, y-col, wt-col, func-to-fit,
                       par-vals, par-consts, fit-ans, tab-opts,
                       tab-results)
```

Argument	Type	I/O	Meaning/Value
fit-text	TEXT	Ouput	Equation of fitted function.
data-tab	TABLE	Input	Observed data in columns of a table
x-col	FIXED or TEXT	Input	Name or index of column of data-tab with independent variable values.
y-col	FIXED or TEXT	Input	Name or index of column of data-tab with observed function values.
wt-col	FIXED or TEXT	Input	Name or index of column of data-tab with weights. If you wish to perform an unweighted fit, this argument must be empty.
func-to-fit	TEXT	Input	The equation of the function you wish to fit to your data. Parameter names must be labeled B1, B2, etc. and the independent variable must be labeled X.
par-vals	FLOAT array	Input	Contains initial values of the parameters such that element i equals the initial value for parameter Bi.

par-consts	TEXT array	Input	Contains any constraints you desire to enforce. If you wish to perform an unconstrained fit, this argument must be empty.
fit-ans	FLOAT array	Output	Upon completion of the fit fit-ans will contain the following statistical results: [1] number of data points [2] multiple R [3] multiple R-squared [4] standard deviation of regression.
tab-opts	FIXED	Input	determines which fit table results will be returned in tab-results. See notes below.
tab-results	TABLE array	Output	Used to obtain results tables. Which tables are returned is determined by the value of tab-opts. See notes below.

Display: none.

Notes

Value of tab-opts	Returned results in tab-results
1	[1] = parameter results table [2] = ANOVA table
2	[3] = table of residuals
3	[1] = parameter results table [2] = ANOVA table [3] = table of residuals
any other value	none

- If your data is in a graph, the 1st argument data-tab must be the points table of the graph.
- If you wish to obtain any graphs of the residuals then you may use the procedure %RESIDUALANALYSIS.
- The syntax of permissable parameter constraints is given in Section 6.
- %FITFUNQT uses MLAB's iterative curve fitting algorithm. Therefore, careful selection of initial parameter values might lead to more accurate fits.

8. FITFUNQT

FITFUNQT quietly performs nonlinear regression on a system of functions in any number of independent variables. It is the quiet entry to the public procedure FITFUN and therefore you must first RAC PUBLIC FITFUN before its use.

FITFUNQT(func-tab, par-tab, data-tab, ind-var-cols, par-vary,
par-consts, fit-ans, tab-opts, tab-results)

Argument	Type	I/O	Meaning/Value
func-tab	TABLE	Input	Used to describe the system you wish to fit to data-tab. See notes below.
par-tab	TABLE	Input	Used to specify the parameters of the system. See notes below.
data-tab	TABLE	Input	Contains independent variable values and observed function data plus any weights (if desired).
ind-var-cols	FIXED array	Input	Specifies the columns of data-tab with values for the independent variables. (The order corresponds exactly to the order in which the independent variables appear in the system reading from left to right and top to bottom.)
par-vary	TEXT array	Input	Contains the names of those parameters which should vary during the fit. (If a parameter does not vary, it will be fixed equal to its initial value throughout the fit.)

par-consts	TEXT array	Input	Contains any constraints you desire to enforce. If you wish to perform an unconstrained fit, this argument must be empty.
fit-ans	FLOAT array array	Output	Upon completion of the fit, the ith row of fit-ans will contain the following statistical results for the ith function in the system. [1] number of data points [2] multiple R [3] multiple R-squared [4] standard deviation of regression.
tab-opts	FIXED	Input	Determines which fit table results will be returned in tab-results. See notes.
tab-results	TABLE array array	Output	Used to obtain results tables. Which tables are returned is determined by the value of tab-opts. See note below.

Display: None.

Notes

Each function included in the system to fit is specified as a row in func-tab which has the following structure:

col 0	col 1	col 2	col 3
TEXT function name	TEXT equation of function	FIXED index of data-tab with observed function values	FIXED index of data-tab with weights for this function (if no weights this entry must be empty).

Each parameter included in the system to fit is specified as a row in par-tab which has the following structure:

col 0	col 1
TEXT parameter name	FLOAT initial value

Value of tab-opts	Returned results in tab-results
1	[1] = parameter results table [2] = ANOVA table
2	[3] = table of residuals
3	[1] = parameter results table [2] = ANOVA table [3] = table of residuals
any other value	none

- Upon return, a new column will be added to your function table which contains the fitted function TEXTs of each function in the system.

- FITFUNQT calls the MLAB curve fitting algorithm which is an iterative procedure. Therefore, careful assignment of initial parameter values might lead to more accurate fits.

- If the system consists of only 1 weighted function, the weights will be normalized. Otherwise, if the system contains several functions, some of which are weighted, the weights will not be normalized.

- If you wish to obtain any graphs of the residuals you may use the procedure %RESIDUALANALYSIS.

9. %FITLINEQT

%FITLINEQT performs simple linear regression, (equivalent to FIT LINE). (See Section 6.)

```
fit-text = %FITLINEQT(data-tab, x-col, y-col, wt-col, forced,
                      fit-ans, tab-opts, tab-results)
```

Argument	Type	I/O	Meaning/Value
fit-text	TEXT	Output	Equation of the fitted line.
data-tab	TABLE	Input	Observed data in columns of the table.
x-col	FIXED or TEXT	Input	Name or index of column of data-tab with independent variable values.
y-col	FIXED or TEXT	Input	Name or index of column of data-tab with observed values.
wt-col	FIXED or TEXT	Input	Name or index of column of data-tab with weights. If you wish to perform an unweighted fit, this argument must be empty.
forced	BOOL	Input	If TRUE the fit will be forced through the origin, otherwise an unforced fit will be performed.
fit-ans	FLOAT array	Output	Upon completion of the fit, fit-ans will contain the following statistical results: [1] number of data points [2] correlation coefficient [3] multiple R-squared [4] standard deviation of regression.

tab-opts	FIXED	Input	Determines which fit table results will be returned in tab-results. See notes.
tab-results	TABLE array	Output	Used to obtain results tables. Which tables are returned and where is determined by the value of tab-opts. See notes.

Display: None.

Notes

Value of tab-opts	Returned results in tab-results
1	[1] = parameter results table [2] = ANOVA table
2	[3] = table of residuals
3	[1] = parameter results table [2] = ANOVA table [3] = table of residuals
any other value	none

- If your data is in a graph, the first argument data-tab must be the points table of the graph.
- If you wish to obtain any graphs of the residuals you may use the procedure %RESIDUALANALYSIS.
- If the fit is exact, no ANOVA table will be returned.

10. %FITPOLQT

%FITPOLQT performs polynomial regression, of fixed degree (equivalent to FIT POLYNOMIAL). (See Section 6.)

```
fit-text =   %FITPOLQT(data-tab, x-col, y-col, wt-col, degree,
                        fit-ans, tab-opts, tab-results)
```

Argument	Type	I/O	Meaning/Value
fit-text	TEXT	Output	Equation of fitted polynomial.
data-tab	TABLE	Input	Contains observed data you wish to fit a polynomial to.
x-col	FIXED or TEXT	Input	Name or index of column of data-tab with independent variable values.
y-col	FIXED or TEXT	Input	Name or index of column of data-tab with observed values.
wt-col	FIXED or TEXT	Input	Name or index of column of data-tab with weights. If you wish to perform an unweighted fit, this argument must be empty.
degree	FIXED	Input	Degree of the polynomial you wish to fit to your data.
fit-ans	FLOAT array	Output	Upon completion of the fit, fit-ans will contain the following statistical results: [1] number of data points [2] multiple R [3] multiple R-squared [4] standard deviation of regression.

tab-opts	FIXED	Input	Determines which fit table results will be returned in tab-results. See notes.
tab-results	TABLE array	Output	Used to obtain the results tables. Which tables are returned and where is determined by the value of tab-opts. See notes.

Display: None.

Notes

Value of tab-opts	Returned results in tab-results
1	[1] = parameter results table [2] = ANOVA table
2	[3] = table of residuals
3	[1] = parameter results table [2] = ANOVA table [3] = table of residuals
any other value	none

- If your data is in a graph, the first argument data-tab must be the points table of the graph.
- If you wish to obtain any graphs of the residuals then you may use the procedure %RESIDUALANALYSIS.
- If the fit is exact, no ANOVA table will be returned.

11. %FRIEDMAN

Performs Friedman's nonparametric tests for blocked data. First performs the overall test of significance and then a multiple comparison test for either all pair-wise comparisons or all comparisons to a control. (See Section 4.)

```
%FRIEDMAN(data,row-col,control-index,all-stats,non-sigs)
```

Argument	Type	I/O	Meaning/Value
tab1	TABLE	Input	Source for data.
row-col	TEXT	Input	"COLUMN" or "C" for columns; "ROW" or "R" for rows (if empty defaults to columns)
control-index	FIXED	Input	Row (column) index for control group (empty if no control group).
all-stats	BOOL	Input	TRUE to obtain all statistics in results table (see Section 4); FALSE results in compact results table.
nonsigs	BOOL	Input	TRUE to obtain results for nonsignificant comparisons; FALSE to obtain only significant results.

Display: Overall results table and multiple comparison results table.

For other multiple comparison tests see %ANOVA, %DOKSUM and %KRUSKALWALLIS.

12. %FSIGLEV

Returns the upper one-tail significance level at X, where X follows an F distribution with df1 and df2 degrees of freedom.

```
level = %FSIGLEV(df1,df2,X)
```

Argument	Type	I/O	Meaning/Value
level	FLOAT	Output	Significance level
df1	FIXED	Input	Degrees of freedom, numerator
df2	FIXED	Input	Degrees of freedom, denominator
X	FLOAT	Input	F-variate

The lower one-tail significance level is 1.- %FSIGLEV (df1,df2,X)

Display: none.

13. %FTEST

Tests for equality of variances of two samples (equivalent to the TEST VARIANCES command). (See Section 3.)

```
%FTEST(data1,data2,df1,df2,F,sig-lev)
```

Argument	Type	I/O	Meaning/Value
data1,data2	FLOAT array	Input	The two sets of data for which variances are to be computed.
df1,df2	FIXED	Output	Degrees of freedom for numerator and denominator respectively.
F	FLOAT	Output	Computed F-statistic.
sig-lev	FLOAT	Output	Significance level of the F-statistic.

Display: none.

For nonparametric test see %DISPTEST.

14. %KRUSKALWALLIS

Performs Kruskal-Wallis nonparametric test for unblocked data. First performs the overall test of significance, then performs a multiple comparison test for either all pairwise comparisons or all comparisons to a control. (See Section 4.)

```
%KRUSKALWALLIS(tab1,row-col,control-index,all-stats,nonsigs)
```

Argument	Type	I/O	Meaning/Value
tab1	TABLE	Input	Source for data.
row-col	TEXT	Input	"COLUMN" or "C" for columns; "ROW" or "R" for rows (if empty defaults to columns)
control-index	FIXED	Input	Row (column) index for control group (empty if no control group).
all-stats	BOOL	Input	TRUE to obtain all statistics in results table; FALSE results in compact results table.
nonsigs	BOOL	Input	TRUE to obtain results for nonsignificant comparisons; FALSE to obtain only significant results.

Display: Overall results table and multiple comparison results table.

For other multiple comparison tests see %ANOVA, %FRIEDMAN, and %DOKSUM.

15. %NORMALITY

Tests an array of data for normality using either the Wilk-Shapiro test when n<=50 or D'Agostino's test when n>50. (See Section 2.)

```
test-result = %NORMALITY(data,stat,sig-lev)
```

Argument	Type	I/O	Meaning/Value
test-result	TEXT	Output	Text of results formatted so it can be displayed.
data	FLOAT array	Input	Data to be tested.
stat	FLOAT	Output	Calculated test statistic.
sig-lev	FLOAT	Output	Significance level of test statistic.

Display: none.

16. %NORMALSHAPE

Tests how a population may deviate from normality by testing whether the skewness and kurtosis of the sample are consistent with the skewness and kurtosis of a normal distribution. This is equivalent to the TEST NORMALSHAPE command. (See Section 2.)

```
test-result = %NORMALSHAPE(data,test-option,stat,sig-lev)
```

Argument	Type	I/O	Meaning/Value
test-result	TEXT	Output	Summary of test results suitable for display.
data	FLOAT array	Input	Data to be tested.
test-option	TEXT	Input	"SKEWNESS" or "S" for test for skewness; "KURTOSIS" or "K" for test for kurtosis.
stat	FLOAT	Output	Computed value of g_1 or g_2.
sig-lev	FLOAT	Output	Significance level of "stat"; tabulated values are used for small sample sizes.

Display: none.

17. %NORMSIGLEV

Returns the two-tail significance level at X, where X follows a normal distribution with mean 0 and variance 1.

```
level = %NORMSIGLEV(X)
```

Argument	Type	I/O	Meaning/Value
level	FLOAT	Output	Significance level
X	FLOAT	Input	Standardized normal deviate.

Display: none.

Note: You can use %TSIGLEV to obtain the cumulative probability density function of the normal distribution or, equivalently, the lower one-tail significance level. In the following example P is the probability of observing a value less than X.

```
IF X<0. THEN P = %NORMSIGLEV(X)/2.;
         ELSE P = 1 - %NORMSIGLEV(X)/2.;
```

The upper one-tail significance level would be 1.-P.

18. %RANKTEST

Performs one- or two-sample nonparametric tests. The Wilcoxon sign rank test is performed if a single sample is compared to a known value or if two samples with paired data are compared. When the data are unpaired the Wilcoxon-Mann-Whitney rank sum test is performed. (See Section 3.)

```
%RANKTEST(data1,data2,test-option,test-result,sig-lev)
```

Argument	Type	I/O	Meaning/Value
data1	FLOAT array	Input	Data for first sample.
data2	FLOAT or FLOAT array	Input	If comparing a single sample it is the standard value, otherwise it is the second data array.
test-option	TEXT	Input	Describes the type of test to be carried out. Values are "VALUE", "PAIRED", "UNPAIRED".
test-result	FLOAT	Output	Resulting test statistic.
sig-lev	FLOAT	Output	Significance level; for small samples tabulated values are used.

Display: Results table.

For parametric test see %TTEST.

19. STATS and MEASURES

To compute STATS and MEASURES, use the commands described in Section 1.

To avoid any questions being asked, you should:

- when adding STATS to a table be sure that there are no measures already added.
- when adding to columns of a table either delete the derivations of derived columns or exclude derived columns using the FOR-clause.

To avoid warnings being typed you should:

- exclude derived columns using the FOR-clause when computing statistics by rows.
- have no zero or negative values when computing LOGSTATS or LOGMEASURES.

You can select the measures to be computed by the STATS and LOGSTATS expressions by preparing a TEXT array to be assigned to %STATSLIST. %STATSLIST can be assigned either a TEXT array variable or a TEXT array constant. It cannot be assigned values element by element. The following examples are correct.

```
DCL MYLIST TEXT ARRAY;
MYLIST[1] = 'MEAN';
MYLIST[2] = 'STDEV';
MYLIST[3] = 'CI95';
%STATSLIST = MYLIST;

%STATSLIST = ['MEAN', 'STDEV', 'CI95'];
```

You can also set the number of decimal places to which a column (or row) of computed values will be rounded off by assigning the FIXED system variable %STATSDP a positive FIXED value.

20. %TFUN

Returns the value of t where the probability of observing a value less than t is p.

```
t = %TFUN(df,p)
```

Argument	Type	I/O	Meaning/Value
t	FLOAT	Output	t-variate
df	FIXED	Input	Degrees of freedom
p	FLOAT	Input	Probability

Display: none.

21. %TSIGLEV

Returns the two-tail significance level at X, where X follows a Student's t-distribution with DF degrees of freedom.

```
level = %TSIGLEV(df,X)
```

Argument	Type	I/O	Meaning/Value
level	FLOAT	Output	Significance level
df	FIXED	Input	Degrees of freedom
X	FLOAT	Input	t-variate

Display: none.

Note: You can use %TSIGLEV to obtain the cumulative probability density function of the Student's t-distribution or, equivalently, the lower one-tail significance level. In the following example P is the probability of observing a t-value less than X.

```
IF X<0. THEN P = %TSIGLEV(DF,X)/2.;
       ELSE P = 1.-%TSIGLEV(DF,X)/2.;
```

The upper one-tail significance level would be 1-P.

22. %TTEST

Performs a t-test to compare the mean of a single normal sample to a known value or to compare the means of two normally distributed samples. (See Section 3.)

```
%TTEST(data1,data2,test-option,df,t,sig-lev)
```

Argument	Type	I/O	Meaning/Value
data1	FLOAT array	Input	Data for the first sample.
data2	FLOAT or FLOAT array	Input	With a single sample--the standard value; otherwise, data array for the second sample.
test-option	TEXT	Input	The type of t-test to be performed. "VALUE" for single sample with known value, "PAIRED" for paired data, "EQUAL" if variances are to be pooled, "UNEQUAL" if variances are unequal.
df	FIXED	Output	Degrees of freedom for the t-statistic.
t	FLOAT	Output	The t-statistic.
sig-lev	FLOAT	Output	Significance level of t-statistic.

Display: Results table of the t-test.

For nonparametric tests see %RANKTEST.

23. %VARIANCETEST

Tests the homogeniety of the variances of more than two samples using either Bartlett's test or Levene's test. (See Section 4.)

```
%VARIANCETEST(table,row-col,test-option,df1,df2,stat,sig-lev)
```

Argument	Type	I/O	Meaning/Value
table	TABLE	Input	Table of data where all numeric data is to be included.
row-col	TEXT	Input	"ROW" or "R" if data organized in rows; "COLUMN" or "C" if data organized in columns.
test-option	TEXT	Input	Options are "BARTLETT" or "B" and "LEVENE" or "L".
df1,df2	FIXED	Output	Degrees of freedom for test statistic; df2 will be empty if test statistic is a chi-squared.
stat	FLOAT	Output	Computed test statistic.
sig-lev	FLOAT	Output	Significance level of the test statistic.

Display: Results table.

Bartlett's test uses a chi-squared statistic and Levene's test uses an F-statistic.

Appendix B: Update to the PROPHET Statistics Manual, March 1985

Section 1: Basic Statistics

1. Public Procedure FDATA2A B-1
2. RANKS expression B-1

Section 2: Distribution Testing

1. %MAKEDISTABLE command B-3
2. Public Procedures CDIST and PSCALE B-3

Section 4: Multisample Hypothesis Testing

1. New Multiple Comparison Procedures B-6
2. %COMPAREORDERED procedure B-9

Section 5: Frequency Analysis

1. CROSSTAB procedure B-10

Section 6: Regression Analysis

1. COVARIANCES and CORRELATIONS expressions B-11
2. %CORRELATION function B-11
3. FIT MULTIPLE command B-11
4. New results from FIT FUNCTION and FITFUN B-13

5. MLAB update B-13
6. ADDFITTEDVALUES procedure B-15

Section 7: Multifactor Analysis of Variance

1. ANOVA procedure B-16

Section 9: Miscellaneous Procedures

1. SURVIVCOMPARE procedure B-30
2. SURVIVGRAPH procedure B-37
3. P2BMDP B-45

Appendix A: Statistical Functions and Procedures

1. New quiet entries for Analysis of Variance, Kruskal-Wallis test and Friedman's test B-49
2. Quiet entries for %JOCKHEEREQT and %PAGEQT .B-50
3. New results from %FITFUNQT and FITFUNQT B-51
4. %FITMULTIPLEQT procedure B-53
5. BLISSPROBIT procedure B-55

Appendix B: Update to the PROPHET Statistics Manual - 3/85

Section 1: Basic Statistics

1. Public Procedure FDATA2A

It is not currently possible to use the PROPHET STATS and MEASURES commands directly to compute descriptive statistics on frequency data in a distribution table. The Public Procedure FDATA2A may be used to "expand" frequency data into a FLOAT array to overcome this.

To use FDATA2A type:

```
RAC PUBLIC FDATA2A  <GO>
float-array = FDATA2A(table)  <GO>
```

PROPHET asks you for the columns containing the data and their associated frequencies.

Following this, you can use any of the usual forms of the STATS and MEASURES expressions on the "float-array".

For example,

```
DARA = FDATA2A(MYTAB)  <GO>
EAD STATS OF DARA  <GO>
SET T TO MEASURES OF DARA  <GO>
```

Note that you do not specify BY ROWS or BY COLUMNS, and you cannot ADD STATS or MEASURES to the FLOAT array specified.

You can also use the resulting array in calls to other statistical procedures that require the input data to be in a FLOAT array.

2. RANKS expression

You can use the RANKS expression to calculate the ordered ranks for data in tables. All tied ranks are averaged. You can compute RANKS for each column independently, for each row independently, or for the table as a whole. The use of RANKS is very similar to that of the STATS and MEASURES expressions. You can use an optional for clause to selectively specify the data you want to use in the analysis. The following examples illustrate the use of RANKS.

```
EAD RANKS OF COLS OF TAB  <GO>
SCAN RANKS OF ROWS OF TAB  <GO>
DISPLAY RANKS OF TAB  <GO>
```

Note that if you do not specify "COLS" or "ROWS" then PROPHET computes the ranks as if the whole table is one sample. For example:

```
PRINT RANKS OF TAB FOR COLS 1 TO 3,6,11,12  <GO>
```

You may also assign the computed ranks of the data table to a temporary table in the usual manner:

```
SET T TO RANKS OF COLS OF TAB FOR COLS <> 1  <GO>
```

Section 2: Distribution Testing

1. %MAKEDISTABLE command

The %MAKEDISTABLE command is used to make distribution tables from data in columns or rows of tables. %MAKEDISTABLE is similar to the MAKE HISTOGRAM command except that it results in a distribution table rather than a graph. The distribution of the data is calculated based on the information that you provide. As in the MAKE HISTOGRAM command dialogue, you are asked for the data table name, whether the data is from rows or columns of the table, and the number or size of intervals. The command to call %MAKEDISTABLE is:

```
CALL %MAKEDISTABLE  <GO>
```

Please refer to the section of this manual on MAKE HISTOGRAM for more information on the distribution of data from PROPHET tables.

2. Public Procedures CDIST and PSCALE

The Public Procedure CDIST is used to calculate the cumulative distribution (percentage) for data in a distribution table. It provides this information by adding a new column to the table. CDIST also computes the cumulative distribution in terms of numbers (probability scale values) that, when plotted, appear as if the cumulative distribution was plotted on "normal probability paper".

To use CDIST type:

```
RAC PUBLIC CDIST  <GO>
CALL CDIST(data-tab)  <GO>
```

where "data-tab" is the name of the distribution TABLE.

An entry to CDIST, CDISTNEW, operates in a similar fashion, except that the data, frequencies, cumulative distribution and probability scale values are put into a new table whose name you specify.

For example,

```
CALL CDISTNEW(data-tab,new-tab-name)  <GO>
```

where "data-tab" is the distribution TABLE and new-tab-name is the TEXT name for the new table.

For example,

```
CALL CDISTNEW(OLDTAB,'NEWTAB')  <GO>
```

Some things to note:

- If the distribution table was made with the MAKE..AS DISTRIBUTION command, the appropriate columns will be selected automatically.

- The distribution table must contain a data column and corresponding frequency column.

- If the distribution table contains more than 2 columns, PROPHET asks for the column numbers of the data and frequency columns.

- If the table already has the cumulative distribution columns, CDIST asks if you wish to continue.

- CDIST does not apply a continuity correction and therefore no entry is made in the "probability-scale" column for the last value in the cumulative distribution.

After calling CDIST or CDISTNEW you may then make a GRAPH of the resultant distribution by using the usual MAKE GRAPH commands and plotting the appropriate columns.

If you plot the "probability" scale values, the vertical tick increments, which are on a linear scale, do not apply. To get a measure of the probability scale on the GRAPH, you can use the Public Procedure PSCALE. PSCALE draws an appropriately labelled vertical line at the right side of the graph (as a curve to the graph).

To use PSCALE type:

```
RAC PUBLIC PSCALE  <GO>
CALL PSCALE(graph)  <GO>
```

where "graph" is the GRAPH to which you want the new "axis" added.

Consider, for example, that you have a table DTAB consisting of two columns, with the distribution data in column 1 and their corresponding frequencies in column 2.

You might then use the following sequence of commands to add the cumulative distribution data to DTAB, make a graph G1 of the percentage distribution, and then make a graph G2 which corresponds to the "normal probability scale" plot.

```
CALL CDIST(DTAB)  <GO>
MAKE GRAPH G1 FROM DTAB AS COL 1 VS COL 33  <GO>

MAKE GRAPH G2 FROM DTAB AS COL 1 VS COL 4  <GO>
CALL PSCALE(G2)  <GO>
```

Appendix B: Update to the PROPHET Statistics Manual - 3/85

Section 4: Multisample Hypothesis Testing

1. New multiple comparison procedures: $MULTICOMP, $DUNCAN, $TUKEY, $SCHEFFE, $DUNNET, $SNK

There are three new multiple comparison procedures which may be used in conjunction with either the 1-way completely randomized ANOVA or the 2-way randomized block ANOVA. These are Tukey's, Duncan's and Scheffe's tests. Each procedure asks for the name of the table, whether the data are organized by rows or columns, and other questions similar to the COMPARE command. Next, the procedures perform the analysis of variance test and the selected multiple comparison test. The format of the results is similar to those presented for the Newman-Keuls test. (See Section IV of the Statistics Manual for examples of the Newman-Keuls test.)

You can use any of these tests with the command:

```
CALL $MULTICOMP  <GO>
```

$MULTICOMP asks you to select the test procedure that you want to use. Alternatively, you can explicitly call any one of the tests outside of $MULTICOMP. For example,

```
CALL $TUKEY  <GO>
CALL $DUNCAN  <GO>
CALL $SCHEFFE  <GO>
```

Note that you can also directly call the procedures for the Newman-Keuls test and Dunnett's test:

```
  CALL $SNK  <GO>
(for Newman-Keuls test)

  CALL $DUNNETT  <GO>
```

$SCHEFFE is somewhat more complicated than the other four procedures because it requires contrast information. A contrast table is required in order to determine which groups are to be compared, and on what basis.

There are three different ways to specify a contrast table for Scheffe's multiple comparison test. These are described below.

i. Performing all pairwise comparisons:

%SCHEFFE asks if you wish to perform all pairwise comparisons. If you answer "YES", all groups are compared to one another, assigning equal weights to each. No contrast table is created in this case.

ii. Making the contrast table:

If you do not want to perform all-pairwise comparisons and you do not have a "template", %SCHEFFE will query you for the information necessary to set up a contrast table. The procedure prompts for the new table's name and asks if you want a permanent copy of the table. The table can then be used as a template for later calls to %SCHEFFE.

You define the contrasts by specifying which groups are to be assigned a positive sign (one "side" of the contrast), and which are to be assigned a negative sign (the other "side" of the contrast). It is possible to assign different weights to each of the groups in the contrast, or to use equal weights.

Next, %SCHEFFE asks for the group numbers (i.e. the row or column numbers) to be used for each side of the contrasts. The numbers of each side are entered as a group with comma or space separators and the list is terminated with <GO>.

If you are not using equal weights, %SCHEFFE asks you to enter the weights for each of the rows/columns specified for each side of the contrast. These are entered in list form as described above, with comma or space delimeters. The sum of the weights for both sides of the contrast must be equal; you can enter any weighting scheme you desire as long as this requirement is met. (Note that all weights are entered as positive numbers. The weights for the second side will automatically be assigned negative values.) If you decide to use fractions, the weights for each contrast side must sum to 1 within a reasonable roundoff. (There is no necessity, for example, of entering a fraction such as 1/3 to more than 2 decimal places.) If a side consists of only one row/column, %SCHEFFE supplies the value of 1. automatically and does not ask you to specify the weights.

When you are done specifying all the contrast information, press <GO> in response to the row/column prompt.

At this point %SCHEFFE displays the contrast table. It contains one row for each contrast and indicates the contrast coefficients for each row/column in the contrast, along with the sums of weighted group means for each side of the contrast.

%SCHEFFE then asks you if you wish to edit the contrast table. If you respond "NO" the program proceeds to perform the analysis of variance and subsequently the Scheffe multiple comparison test. If you respond "YES" you can then respecify portions of the contrast table. PROPHET first asks for the contrast you wish to edit, and then prompts for the information needed to respecify the contrast (as it did when you made the table originally). When you have finished making the first change %SCHEFFE asks for the next contrast table you want to edit. If you just press <GO> PROPHET displays the table again, and you will have the option of making further changes by responding "YES" to the edit question. If you respond "NO" the rest of the procedure will proceed as described above.

iii. Using a "template" contrast table:

There may be situations where you have a similar experimental design for several different groups of observations. In these cases you might want to make a permanent copy of the contrast table that can then be used as a model for subsequent uses of the Scheffe test; in this way you would only have to make the table once.

%SCHEFFE first asks if you have a "template" table to use. If you respond "YES" , it asks for the name of the table.

Note that data sets used with template contrast tables must have the same number of groups.

2. %COMPAREORDERED procedure

%COMPAREORDERED performs a nonparametric comparison of more than two groups when the groups are ordered. That is, the null hypothesis is that the measures of central tendency are all equal and the alternate hypothesis is that they have a specific order. For example, you might have divided your subjects according to their health: good, fair, and poor. If the data are unblocked, PROPHET performs the Jonckheere-Terpstra test, otherwise it performs Page's test. Note that the corresponding tests in COMPARE, Kruskal-Wallis and Freedman's tests, are based on the alternative hypothesis that the groups are different but not necessarily in any order. To perform either test with ordered alternatives type:

```
%COMPAREORDERED  <GO>
```

PROPHET then prompts you for the remaining information in the same manner as COMPARE does when you are comparing more that two groups. PROPHET also asks you:

```
Are the groups ordered from smallest to largest?
```

If you answer YES then PROPHET uses the data in the same order as in your data table. If you answer NO then PROPHET asks you:

```
List row (column) numbers of groups in order, separated
by spaces or commas.
```

PROPHET then performs Page's test or Jonckheere-Terpstra test depending on whether you said the data are blocked or not blocked, respectively.

Section 5: Frequency Analysis

1. CROSSTAB revision

The Public Procedure CROSSTAB has been revised so that the input data table need not contain only integer data. Now you can use either numeric or textual data that can be categorized in the following ways:

by integer values,
within numeric ranges, or
by text values.

When a data column is numeric CROSSTAB gives you the option to use ranges or integer values. If you choose to use ranges it asks you to specify the points which divide the data into groups (cutpoints). For example you could divide the data into groups in the following way:

```
COLUMN NUMBER OF THE 1ST VARIABLE (ROW VAR.)?  1 <GO>
DO YOU WANT TO SPECIFY CUTPOINTS FOR THE ROW VARIABLE?
          YES    <GO>
ENTER CUTPOINTS SEPARATED BY COMMAS OR SPACES:
          0,2,3  <GO>
```

These three cutpoints divide the data into four groups:

- less than or equal to 0
- greater than 0 but less than or equal to 2
- greater than 2 but less than or equal to 3
- greater than 3

If you do not use cutpoints the data are divided into groups according to their (rounded) integer values. When a column contains text values you have the option of using all the values found in the column or specifying those values you would like to include. The groups in the contingency table will be arranged in the same order as the specified list, the column's descriptor list (if it has one), or the order of occurrence in the column. For example:

```
COLUMN NUMBER OF THE 2ND VARIABLE (COL VAR.)?  2  <GO>
DO YOU WANT TO SPECIFY A SET OF VALUES
             FOR THE COLUMN VARIABLE?  YES  <GO>
SPECIFY TEXT VALUES SEPARATED BY SPACES
             OR COMMAS:  YELLOW,RED,BLUE    <GO>
```

Section 6: Regression Analysis

1. COVARIANCES and CORRELATIONS expressions:

The CORRELATIONS and COVARIANCES expressions allow you to compute either the correlation or covariance matrix of the rows or columns of a TABLE (or 2-dimensional ARRAY). The form of the commands is similar to the STATS and MEASURES commands. Some examples are:

```
DISPLAY CORRELATIONS OF COLUMNS OF DATATAB   <GO>
SCAN COVARIANCES OF ROWS OF DATATAB   <GO>
SET T TO CORRELATIONS OF DATATAB FOR COLS <> 1   <GO>
```

In the first example the Ith row and Jth column of the resulting table contains the correlation of columns I and J of DATATAB.

The results are returned in a temporary table which can be viewed using any of the display commands (e.g., EAD, DISPLAY, SCAN, PRINT, PRINTOUT) or named using the SET command. If you do not specify whether to use the rows or the columns of the data table, PROPHET assumes columns. You can also use "FOR clauses" to specify the column or row to include (or exclude). PHOPHET excludes nonnumeric columns automatically.

PROPHET automatically rounds correlations to 3 decimal places, regardless of the value you may have set in the SELECT STATS quiz.

2. %CORRELATION Function

The function %CORRELATION takes as arguments two FIXED or FLOAT arrays and returns the correlation coefficient of the two groups. For example,

```
TYPE %CORRELATION(ARA1,ARA2)  <GO>
R = %CORRELATION(D1,D2)  <GO>
```

3. FIT MULTIPLE

The FIT MULTIPLE command performs stepwise multiple linear regression. You use it in a manner analogous to the other fitting programs (e.g., FIT LINE) in PROPHET.

NOTE: The FIT MULTIPLE command totally supersedes the Public Procedures MULREG and MREG.

You can use the FIT MULTIPLE command in 4 different ways:

i. FIT MULTIPLE <GO>

PROPHET asks you to supply the name of your DATA TABLE, the column with Y DATA and the columns with X DATA.

ii. FIT MULTIPLE TO table <GO>

In this case, "table" is the name of your DATA TABLE. PROPHET assumes that the last column of "table" contains the Y data and all other columns contain the X data.

iii. FIT MULTIPLE TO COLUMN y-col OF table <GO>

In this case, y-col is either the FIXED column number or TEXT column heading of your data table which contains the Y DATA. PROPHET assumes that all other columns contain X DATA.

iv. FIT MULTIPLE TO COLUMN y-col OF table FOR
COLUMNS = x-cols <GO>

This case is identical to case iii except that you use a FOR CLAUSE to specify which columns of the data table contain X DATA. For example:

```
FIT MULTIPLE TO COLUMN 6 OF
DTAB FOR COLS=1 TO 3,5,7 <GO>
```

After you give any of the above commands, PROPHET gives the option of seeing the table of correlations for your specified data. Next, you will enter an interactive loop which performs stepwise multiple regression governed by your directions. Prior to adding or deleting terms, PROPHET shows you a TABLE OF VARIABLES NOT IN THE MODEL that includes useful information describing how the addition of each available variable to enter will affect the model. PROPHET suggests the next best next action (in terms of maximizing the F-ratio you should take). You can then specify the X-variables that you want to enter and/or remove.

After a fit is performed, PROPHET presents a text summary plus a parameter results table and an ANOVA table. Next, it shows the table of variables that are not in the model again, and the interaction proceeds as above.

You terminate the stepwise regression by typing <GO> in response to prompts for variables to enter or to remove. At this point PROPHET asks whether you would like to see a table of residuals corresponding to your final model. If desired, you may also examine graphs of the residuals.

Missing values in FIT MULTIPLE are handled in the following way. If any observation (i.e. row) in the table contains an empty cell in any of the columns you specified for your set of x-variables, that entire row is excluded from the analysis. If you subsequently decide that the x-variable (i.e., column) containing the empty cell is not important to your model, you should then call FIT MULTIPLE again, this time excluding that column by using the FOR CLAUSE as described above.

The meaning of the various terms in the table of variables not in the model can be determined from any standard statistics text (see for example, Draper and Smith, "Applied Regression Analysis"). The TOLERANCE term is a measure of the contribution the corresponding variable will have given variables already in the model. In other words, it describes the extent to which that variable is linearly dependent on other terms already in the model. If the tolerance is LOW, adding it to the model may lead to singularity of the resultant matrix, i.e., it should not be added. Conversely, if the value of the tolerance is close to 1, its contribution is not explained by variables already in the model, it will not be likely to lead to singularity of the matrix, and it can be incorporated into the model.

4. Additional results from FIT FUNCTION and FITFUN:

When you use the procedures FIT FUNCTION and FITFUN after MLAB has completed a fit, PROPHET reports the number of iterations the curve-fitter required along with a statement stating whether or not it converged. Also, if you enforced any constraints, PROPHET displays a constraints table that contains the associated Lagrange multipliers.

5. MLAB update:

The major new features of MLAB include:

i. The inclusion of CLAB operators for cluster analysis

ii. Enhanced graphics and character fonts

iii. Extension of the FIT command including the employment of norms besides the 2-norm.

iv. The new "PROPHET" command which, when given in MLAB, will save your MLAB workarea and transfer control back to your PROPHET session. Subsequently, when you run MLAB again, PROPHET will reinstate the previously saved MLAB workarea - exactly as you left it. The new "PROPHET" command in addition to the PROPHET PUBLIC procedures "P2MLAB" and "MLAB2P" provide the capability of using both systems in concert to analyze a problem.

A comprehensive scenario includes the following:

a. Using the PROPHET PUBLIC procedure "P2MLAB" to save PROPHET data in an MLAB save file and then running MLAB.

b. Using MLAB.

c. Saving your intermediate MLAB results in an MLAB save file.

d. Giving the "PROPHET" command to save your MLAB workarea and return to PROPHET

e. Using the PROPHET PUBLIC procedure "MLAB2P" to retrieve these MLAB results into PROPHET tables

f. Giving the "MLAB" command (enabled by calling P2MLAB) to return to MLAB and resume in the previously saved workarea.

v. Of particular interest is the capability of specifying what version of MLAB to use. Currently, the default version is named "MLAB" and is the newest version which supports the "PROPHET" command. If you wish to use the previous version (named "OLDLAB", the default version in PROPHET 65) you need only do:

```
$MLABVERSION = 'OLDLAB'  <GO>
```

Subsequently, to use the system default version (the newer version), do:

```
$MLABVERSION = $SYSTEMMLABVERSION  <GO> or
      $MLABVERSION = 'MLAB'  <GO>
```

6. %ADDFITTEDVALUES

You can use the procedure %ADDFITTEDVALUES to add (or insert) a derived column to calculate predicted values from a fitted function. The form of the command is:

```
CALL %ADDFITTEDVALUES(tab,x-col,fit-text,fit-col)  <GO>
```

where:

tab	is a table of data
x-col	is a FIXED specifying the column number for the x-data, or a text of the column heading
fit-text	is a text specifying the function in terms of "x" such as as fitted function, returned by any of the FIT commands
fit-col	is a FIXED specifying the column number to insert for the fitted data. If fit-col is empty, the fitted values will be added as a new column to the table.

Section 7: Multifactor Analysis of Variance

1. ANOVA

by Margaret Reid-Miller
Bolt Beranek and Newman Inc.

ANOVA performs analysis of variance in PROPHET for any model of two to six factors with equal number of replications. Models include a wide variety of fixed and random effect models, including complete factorial, randomized blocks, repeated measures and nested designs. ANOVA is flexible enough to allow your data to be organized in tables in almost any way you choose.

To call ANOVA, type:

```
RAC PUBLIC ANOVA <GO>
CALL ANOVA <GO>
```

ANOVA first asks you a number of questions to determine how your data are organized. Or you may use a "design" table that you made during a previous call to ANOVA to specify the data. Next, ANOVA asks you to select the analysis of variance model you would like to use from a list of models with the appropriate numbers of factors. You have the option of seeing an example of any model in the list. The examples are somewhat artificial to allow easy comparison among models. Finally, ANOVA shows the Analysis of Variance table for the model you chose. The use of ANOVA is illustrated through several examples.

Example 1: Two factor analysis of variance with replication (Zar p. 163)

Figure 1-1 shows the table CALCIUM which contains calcium levels of birds under several conditions. There are two factors simultaneously being tested. The first factor is hormone treatment (HORMONE) with two levels (hormone treated and not treated) and the second factor is sex (SEX) also with two levels (male and female). For each of the four combinations of hormone and sex there are five birds.

CALCIUM: Plasma Calcium Levels (mg/100 ml) of Birds

ROWNAME	No Hormone Female	No Hormone Male	Hormone Female	Hormone Male
1.	16.5	14.5	39.1	32.
2.	18.4	11.	26.2	23.8
3.	12.7	10.8	21.3	28.8
4.	14.	14.3	35.8	25.
5.	12.8	10.	40.2	29.3
6. -N-	5.	5.	5.	5.
7. -SUM-	74.4	60.6	162.6	138.9
8. -MEAN-	14.88	12.12	32.52	27.78
9. -VARIANCE-	6.216995	4.476997	69.717	11.182
10. -STDEV-	2.49339	2.115892	8.34967	3.34395
11. -CI95-	3.099917	2.63059	10.38076	4.157378

Figure 1-1 Table of two factor experiment

Specifying Data in Standard Order

Figure 1-2 shows the ANOVA dialogue for specifying your data organization. The first question asks you to select one of four options on how you would like to specify your data. In this example, option 1 for data in "standard" order is the best choice (a more complete description of what is meant by standard order is given below under the heading "Standard Order".) ANOVA asks for the number of factors and the name and number of levels for each factor. (You need not specify the replications factor as required by the public procedure ARAFIL, now superceded by ANOVA). Then ANOVA asks whether the data are in the same table and for the name of the table. Finally, ANOVA asks whether different levels are in different rows or columns for each factor.

```
How do you want to specify your data:

1) make a 'design' table for data in 'standard' order;
2) make a design table for data NOT in standard order;
3) use an existing design table; or
4) use classification variables for each factor?1←

Number of factors?2←
Name for factor 1?HORMONE← How many levels are in factor HORMONE?2←

Name for factor 2?SEX← How many levels are in factor SEX?2←

Are data for all levels of factor HORMONE in the same data table?Y←
  Name of table?CALCIUM←
Are data for different levels of factor HORMONE
  in different rows (R) or columns (C)?C←
Are data for different levels of factor SEX
  in different rows (R) or columns (C)?C←
```

Figure 1-2 ANOVA dialogue specifying the data

Design Table

At this point, ANOVA asks whether you would like to see the design table and whether you would like it in a temporary table for later use, as shown in Figure 1-3. Each row of the design table specifies one column of a data table. For example, the second row indicates that the data for first level of HORMONE (treated) and the second level of SEX has 5 replications and the data are found in the table CALCIUM, in column 2, from rows 1 to 5. Notice that ANOVA checks whether STATS or MEASURES were added to the data table and ignores those rows or columns.

```
Would you like to see the design table?Y

  4R X 7C
 |  | HORMONE | SEX | # of         | Table   | Col | First | Last |
 |  | (H)     | (S) | Replications | Name    |  #  | Row   | Row  |
 |  |         |     | (R)          |         |     |       |      |
 |1 | 1       | 1   | 5            | CALCIUM | 1   | 1     | 5    |
 |2 | 1       | 2   | 5            | CALCIUM | 2   | 1     | 5    |
 |3 | 2       | 1   | 5            | CALCIUM | 3   | 1     | 5    |
 |4 | 2       | 2   | 5            | CALCIUM | 4   | 1     | 5    |

Would you like the design data in a TEMPORARY table?
```

Figure 1-3 Design table for CALCIUM

Model Selection

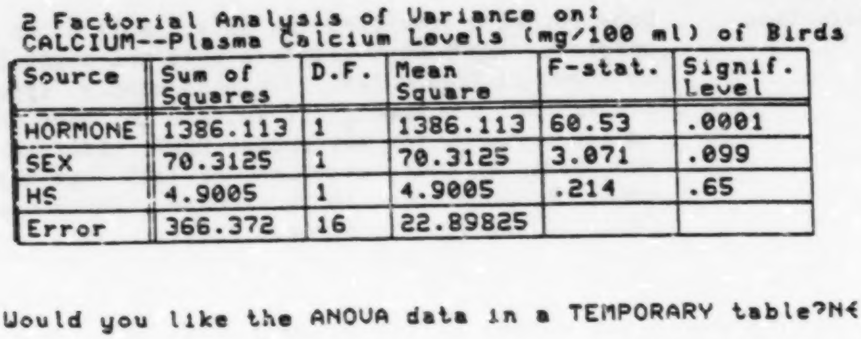
Once you have specified the data, ANOVA shows a list of possible experimental models that you might have used, as shown in Figure 1-4. You need simply type in the number of the model. If you are not sure which model is appropriate, you may choose HELP. HELP asks whether you want to see examples of all the models or an example of one model in particular. In this example, 2 factorial is the appropriate model.

```
Please enter the number of the experimental model you used:
1) 2 Factorial
2) 2 Factorial, blocked
3) 3 Factorial, no replications
4) 1 Repeated Measure
5) 2 Repeated Measures
6) Nested
7) Help?7←

Which model do you want to see an example of (type CTRL-E for all
  models)?1←

2 Factorial:
e.g., There are two grouping factors: 2 levels of HYPERTENSION [H]
(with and without) and 3 levels of DOSAGE [D] (low, medium and high).
Equal numbers of rats are randomly assigned to the 6 combinations
of hypertension and dosage.
MODEL:  H, D, R(HD) -- R random

Please enter the number of the experimental model you used:
1) 2 Factorial
2) 2 Factorial, blocked
3) 3 Factorial, no replications
4) 1 Repeated Measure
5) 2 Repeated Measures
6) Nested
7) Help?1←
```

Figure 1-4 Model selection

Analysis of Variance Table

Figure 1-5 shows the resulting analysis of variance table. Included are the main effects for HORMONE treatment and SEX, the HORMONE by SEX interaction and the ERROR term. Except for the main effects, ANOVA uses one letter abbreviations for the factor names. These abbreviations are shown in parentheses in the factor column readings of the design table (see Figure 1-3). For each effect, the sum of squares, the degrees of freedom, mean square, F-statistic and the significance level are given. Finally, ANOVA asks whether you would like to have the analysis of variance data in a temporary table.



2 Factorial Analysis of Variance on:
CALCIUM--Plasma Calcium Levels (mg/100 ml) of Birds

Source	Sum of Squares	D.F.	Mean Square	F-stat.	Signif. Level
HORMONE	1386.113	1	1386.113	60.53	.0001
SEX	70.3125	1	70.3125	3.071	.099
HS	4.9005	1	4.9005	.214	.65
Error	366.372	16	22.89825		

```
Would you like the ANOVA data in a TEMPORARY table?N<
```

Figure 1-5 Analysis of variance table

Data Specification Options

ANOVA lets you specify how your data are organized in four different ways (see Figure 1-2). The first option is for when your data are in "standard" order. For example, the data in Figure 1-1 are in standard order. You would use option 2 when your data are not in standard order. For example, if columns 1 and 2 of CALCIUM were interchanged, then the data are not in standard order. Option 3 lets you use a design table that was made by a previous call to ANOVA. (You can also use design tables that were made by the public procedure ARAFIL.) The design table specifies how your data are organized. Finally, you may choose to use option 4 to specify classification variables which specify the factor combination for each observation. The following examples show how to use these options.

Example 2: Nested analysis of variance (Zar p. 144)

Figure 2-1 shows the blood cholesterol concentrations of women receiving three different drugs (DRUGS) where each drug has been obtained from two different sources (SOURCES) and the sources differ for each drug. In this example, SOURCE is nested under DRUG because the unique effects of sources A and Q on cholesterol levels are confined to drug 1, whereas the unique effects of sources D and B are confined to drug 2 and sources L and S to drug 3.

COLEST: Blood Cholesterol Levels in Women

0 ROWNAME	1 DRUG 1 SOURCE A	2 DRUG 1 SOURCE Q	3 DRUG 2 SOURCE D	4 DRUG 2 SOURCE B	5 DRUG 3 SOURCE L	6 DRUG 3 SOURCE S
1. REP 1	102.	103.	108.	109.	104.	105.
2. REP 2	104.	104.	110.	108.	106.	107.

Figure 2-1 Table of a nested design

Figure 2-2 shows the design table that was made in a previous run of ANOVA for the data in Figure 2-1. The data organization is analogous to the organization in Example 1 except that the first factor has three levels instead of two. Figure 2-3 shows the corresponding dialogue for specifying the data using the design table.

6R X 7C

	1 DRUG (D)	2 SOURCE (S)	3 # of Replications (R)	4 Table Name	5 Col #	6 First Row	7 Last Row
1	1	1	2	COLEST	1	1	2
2	1	2	2	COLEST	2	1	2
3	2	1	2	COLEST	3	1	2
4	2	2	2	COLEST	4	1	2
5	3	1	2	COLEST	5	1	2
6	3	2	2	COLEST	6	1	2

Figure 2-2 Design table for data in Figure 2-1

```
How do you want to specify your data:

1) make a "design" table for data in "standard" order;
2) make a design table for data NOT in standard order;
3) use an existing design table; or
4) use classification variables for each factor?3<

Name of design table?COLESTDESIGN<
```

Figure 2-3 ANOVA dialogue using a design table

Nested Designs

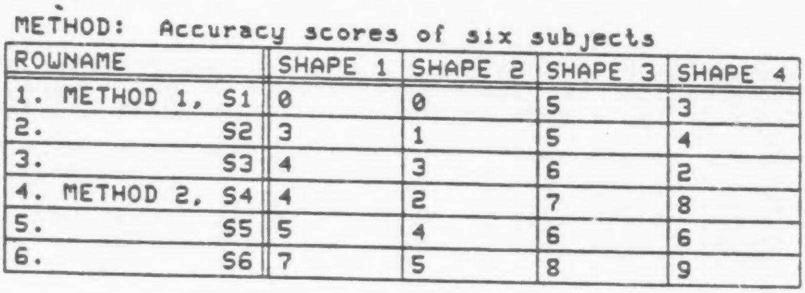
For this type of analysis, you should choose the "nested" model, as shown in Figure 2-4. Note that ANOVA assumes that the second factor, SOURCE, is nested within the first factor, DRUG. There are two main effects, DRUG and SOURCE(D), i.e., SOURCE within DRUG. Each has their own error term shown below the main effect. (For nested designs, the error term for DRUG is SOURCE(D) itself. Therefore, the DRUG's error term and SOURCE(D) have the same sum of squares and degrees of freedom). No interaction term is inlcuded because it is not possible to evaluate the interaction of DRUG and SOURCE when each SOURCE is used for one drug only.

```
Please enter the number of the experimental model you used:
1) 2 Factorial
2) 2 Factorial, blocked
3) 3 Factorial, no replications
4) 1 Repeated Measure
5) 2 Repeated Measures
6) Nested
7) Help?6<


Nested Analysis of Variance on:
COLEST--Blood Cholesterol Levels in
Women (Zar p. 144)
+-----------+----------+------+----------+---------+--------+
| Source    | Sum of   | D.F. | Mean     | F-stat. | Signif.|
|           | Squares  |      | Square   |         | Level  |
+-----------+----------+------+----------+---------+--------+
| DRUG      | 61.16667 | 2    | 30.58333 | 61.17   | .004   |
+-----------+----------+------+----------+---------+--------+
| Error     | 1.5      | 3    | .5000001 |         |        |
+-----------+----------+------+----------+---------+--------+
| SOURCE(D) | 1.5      | 3    | .5000001 | .3333   | .802   |
+-----------+----------+------+----------+---------+--------+
| Error     | 9.       | 6    | 1.5      |         |        |
+-----------+----------+------+----------+---------+--------+

Would you like the ANOVA data in a TEMPORARY table?N<
```

Figure 2-4 Nested analysis of variance

Example 3: One grouping and one repeated measure factors (Winer p. 525)

Figure 3-1 shows a table where each row corresponds to a different subject. Each subject was asked to calibrate four different shaped dials (SHAPE). The order the subject used the dials was randomized. The subjects were divided into two groups, each group using a different calibration method (METHOD). In this example, METHOD is a grouping factor because each subject uses one method only and subjects are grouped according to the methods they use, while SHAPE is a repeated measure factor because each subject repeats the experiment for all levels of SHAPE.



METHOD: Accuracy scores of six subjects

ROWNAME	SHAPE 1	SHAPE 2	SHAPE 3	SHAPE 4
1. METHOD 1, S1	0	0	5	3
2. S2	3	1	5	4
3. S3	4	3	6	2
4. METHOD 2, S4	4	2	7	8
5. S5	5	4	6	6
6. S6	7	5	8	9

Figure 3-1 Table of one repeated measure experiment

Specifying Data Not In Standard Order

Although these data are in standard order, Figure 3-2 shows the dialogue for entering data not in standard order. As before, ANOVA asks for the number of factors and then the names and number of levels for each factor. NOTE: For repeated measures designs, you must specify the grouping factor (METHOD) before the repeated measure factor (SHAPE). Next, ANOVA asks for the number of replications and whether the replications are in rows or columns. The replications (SUBJECTS) for a single dial shape and method are in different rows.

```
How do you want to specify your data:

1) make a "design" table for data in "standard" order;
2) make a design table for data NOT in standard order;
3) use an existing design table; or
4) use classification variables for each factor?2<

Number of factors?2<
Name for factor 1?METHOD< How many levels are in factor METHOD?2<

Name for factor 2?SHAPE< How many levels are in factor SHAPE?4<
Are there replications also?Y<
Number of replications?3<
Are replications for a single factor combination in different
rows (R) or columns (C)?R<

Please enter information for the following factor combinations:
 METHOD1 SHAPE1:  Table name?METHOD< Col $?1< First Row?1<
 M1 S2:  Table name [Press <GO> if METHOD]?< Col $?2< First Row?1<

 M1 S3:  Table name [METHOD]?< Col $?3< First Row?1<
 M1 S4:  Table name [METHOD]?< Col $?4< First Row?1<
 M2 S1:  Table name [METHOD]?< Col $?1< First Row?4<
 M2 S2:  Table name [METHOD]?< Col $?2< First Row?5<
 M2 S3:  Table name [METHOD]?< Col $?3< First Row?4<
 M2 S4:  Table name [METHOD]?< Col $?4< First Row?4<
```

Figure 3-2 Dialogue for data not in standard order

For every factor combination, ANOVA asks for the name of the data table, column number for the combination, and the row of the first observation. ANOVA specifies the factor combination by using the single letter factor abbreviations and the level index. For example, in Figure 3-2 "M2 S1" indicates the second level of METHOD and the first level SHAPE. You need only press <GO> in response to the table name prompt if the table name is the same as the previous factor combination.

Editing the Design Table

Then ANOVA shows you the completed design table and gives you an opportunity to correct any errors. Figure 3-3 shows the resulting design table and the user correcting row 6. ANOVA asks you to reenter the table name, column number, and first row number. The current value for each of these items is shown in parenthesis. ANOVA continues to give you the option to correct rows until you are done. When you press CRTL-E in response to the row number to edit ANOVA shows the design table again and asks whether you want to edit the table again. Finally, it asks whether you would like the design data in a temporary table.

```
Do you want to edit the design table?Y<
```

8R X 7C

	METHOD (M)	SHAPE (S)	$ of Replications (R)	Table Name	Col $	First Row	Last Row
1	1	1	3	METHOD	1	1	3
2	1	2	3	METHOD	2	1	3
3	1	3	3	METHOD	3	1	3
4	1	4	3	METHOD	4	1	3
5	2	1	3	METHOD	1	4	6
6	2	2	3	METHOD	2	5	7
7	2	3	3	METHOD	3	4	6
8	2	4	3	METHOD	4	4	6

```
Row of design table to edit?6< M2 S2 Table name (METHOD)?< Col  $
   (2)?2< First Row (5)?4<

Row of design table to edit.  (Enter <CTRL-E> when you are done or
when you would like to scan the design table again)?<Empty><
```

8R X 7C

	METHOD (M)	SHAPE (S)	$ of Replications (R)	Table Name	Col $	First Row	Last Row
1	1	1	3	METHOD	1	1	3
2	1	2	3	METHOD	2	1	3
3	1	3	3	METHOD	3	1	3
4	1	4	3	METHOD	4	1	3
5	2	1	3	METHOD	1	4	6
6	2	2	3	METHOD	2	4	6
7	2	3	3	METHOD	3	4	6
8	2	4	3	METHOD	4	4	6

```
Do you want to edit the design table again?N<
Do you want the design data in a TEMPORARY table?N<
```

Figure 3-3 Editing the design table

The model selection and final analysis of variance table for one repeated measure design is shown in Figure 3-4. In this design the grouping factor METHOD has its own error term while the SHAPE effect and the METHOD by SHAPE interaction have a different error term.

```
Please enter the number of the experimental model you used:
1) 2 Factorial
2) 2 Factorial, blocked
3) 3 Factorial, no replications
4) 1 Repeated Measure
5) 2 Repeated Measures
6) Nested
7) Help?4<


 1 Repeated Measure Analysis of Variance on:
 METHOD--Accuracy scores of six subjects
 +--------+-----------+------+-----------+---------+---------+
 | Source | Sum of    | D.F. | Mean      | F-stat. | Signif. |
 |        | Squares   |      | Square    |         | Level   |
 +--------+-----------+------+-----------+---------+---------+
 | METHOD | 51.04167  | 1    | 51.04167  | 11.89   | .026    |
 | Error  | 17.16666  | 4    | 4.291666  |         |         |
 | SHAPE  | 47.45833  | 3    | 15.81944  | 12.8    | .0005   |
 | MS     | 7.458333  | 3    | 2.486111  | 2.011   | .166    |
 | Error  | 14.83333  | 12   | 1.236111  |         |         |
 +--------+-----------+------+-----------+---------+---------+

Would you like the ANOVA data in a TEMPORARY table?N<
```

Figure 3-4 Repeated measure analysis of variance

Standard Order

As the previous examples show, your data can be in a variety of different formats and still be in standard order. The restrictions are:

i) either all the levels for a factor must be in the same table or each level must be in a different table

ii) Factor combinations within a table must be in lexagraphical order (see below for an explanation)

iii) Replications must be all in columns or all in rows and there must be the same number of rows or columns for each factor combination

Let Yijk represent the observations for the ith level of factor 1, the jth level of factor 2 and the kth level of factor 3. If the data are in columns, then the lexagraphical ordering of the columns would be:

Y111 Y112 Y121 Y122 Y211 Y212 Y221 Y222

The index corresponding to the first factor varies more slowly than the index to the second factor, and the index to the second factor varies more slowly than the index of the third factor. Notice that all levels of the second factor appear as ordered blocks within each level of the first factor and all levels of the third factor appear in order within each factor combination of the first two factors.

Example 4. Two factors using classification variables

Figure 4-1 shows the data in Figure 1-1 reorganized so that all the observations are all in column 3 of the table BIRDS. In Figure 1-1 the levels of the factors corresponding to an observation are implicitly defined by the observation's row and column position in CALCIUM. In Figure 4-1 the levels of the factors corresponding to an observation are explicitly given in the first two columns, TREATMENT and SEX. These two columns are known as classification variables.

BIRDS: Plasma Calcium Levels (mg/100 ml) of Birds

	1 TREATMENT	2 SEX	3 CALCIUM
1	NO HORMONE	FEMALE	16.5
2	NO HORMONE	FEMALE	18.4
3	NO HORMONE	FEMALE	12.7
4	NO HORMONE	FEMALE	14.
5	NO HORMONE	FEMALE	12.8
6	NO HORMONE	MALE	14.5
7	NO HORMONE	MALE	11.
8	NO HORMONE	MALE	10.8
9	NO HORMONE	MALE	14.3
10	NO HORMONE	MALE	10.
11	HORMONE	FEMALE	39.1
12	HORMONE	FEMALE	26.2
13	HORMONE	FEMALE	21.3
14	HORMONE	FEMALE	35.8
15	HORMONE	FEMALE	40.2
16	HORMONE	MALE	32.
17	HORMONE	MALE	23.8
18	HORMONE	MALE	28.8
19	HORMONE	MALE	25.
20	HORMONE	MALE	29.3

Figure 4-1 Table with classification variables

Specifying Classification Variables

For this data, therefore, option 4 should be used to specify the data, as shown in Figure 4-2. ANOVA first asks for the number of factors. Then it asks for the column number of the classification variables. When the column is of type TEXT, as the two columns in this example are, ANOVA asks whether you want to specify the values to include. If you answer NO then each value in the column forms its own classification level. If you answer YES then you would list the values to include separated by spaces or commas. You can also include "OTHER" in this list. If you specify OTHER then all values not in the list would be included in the single classification level OTHER.

```
How do you want to specify your data:

1) make a 'design' table for data in 'standard' order;
2) make a design table for data NOT in standard order;
3) use an existing design table; or
4) use classification variables for each factor?4<

Name of table with classifications?BIRDS< Number of classification
   variables?2<
Column number with classifications for variable 1?1< Do you want
   to specify the values to include?N<
Column number with classifications for variable 2?2< Do you want
   to specify the values to include?N< Column number with reponses?3<
   Would you like to see the reformatted data?Y<
```

Figure 4-2 ANOVA dialogue for classification variables

ANOVA asks whether you would like to see the data reformatted so that they are in standard order and if you would like to have the data in a TEMPORARY table, as shown in Figure 4-3. With the data in standard order you could later perform other statistical analyses of the data on PROPHET. From this point on ANOVA does the same things as in Example 1, see Figures 1-3, 1-4 and 1-5.

```
Column 'CALCIUM' of BIRDS--Plasma Calcium Levels (mg/100 ml) of Bird
   5R X 4C
+---++--------+--------+-----------+-----------+
|   || 1      | 2      | 3         | 4         |
|   || HORMONE| HORMONE| NO HORMONE| NO HORMONE|
|   || FEMALE | MALE   | FEMALE    | MALE      |
+---++--------+--------+-----------+-----------+
| 1 || 39.1   | 32.    | 16.5      | 14.5      |
| 2 || 26.2   | 23.8   | 18.4      | 11.       |
| 3 || 21.3   | 28.8   | 12.7      | 10.8      |
| 4 || 35.8   | 25.    | 14.       | 14.3      |
| 5 || 40.2   | 29.3   | 12.8      | 10.       |
+---++--------+--------+-----------+-----------+

Would you like the reformatted data in a TEMPORARY table?N<
```

Figure 4-3 The data in Figure 4-1 reformatted

Section 9: Miscellaneous Procedures

1. SURVIVCOMPARE

by Margaret Reid-Miller
Bolt Beranek and Newman Inc.

Summary

SURVIVCOMPARE compares two or more stratified or unstratified cumulative survival curves, where the data are time periods from one well defined event to another well defined event.

Calling Sequence

```
RAC PUBLIC SURVIVCOMPARE <GO>
CALL SURVIVCOMPARE <GO>
```

Entry

```
CALL SURVIVCOMPAREQT (data-table,       event-cols,
                       <table> or     <fixed array> or
                      <text array>   <fixed array array>

                       censored-cols,   mgtest, results-tab,
                     <fixed array> or   <text>    <table>
                    <fixed array array>

                       covar-tab,   df,    chisq,  siglev)
                        <table>  <fixed>  <float>  <float>
```

Questions or Comments

Call the PROPHET Hotline.

Things to Remember

For each group, event and censored times must be in separate columns.

Introduction

SURVIVCOMPARE compares two or more cumulative survival curves where the data are time periods from one well defined event to another well defined event. For example, you might observe the duration from start of a treatment until relapse or death due to disease. If the time to the actual event is unknown (e.g., lost to follow-up, accidental death, or event after end of study) then the last known time is recorded as a censored time.

The two available tests, Mantel-Cox log-rank test and Gehan-Breslow test, are analogues of the nonparametric rank tests. The Mantel-Cox log-rank test tends to give more weight to the later times, and so will obtain significance if the curves are separate at later times. Whereas the Gehan-Breslow test weighs the earlier times more heavily and will note an early separation even if the curves eventually come together or cross. Different results from the two tests may indicate an interaction between time and treatment.

Data

SURVIVCOMPARE expects the data for all the groups in one table with times to events and times up to censoring in separate columns. The columns can be in any order and you need not have both an event column and a censored column for each group. However, you can use a "standard" ordering (see below) so that you do not need to specify each column explicitly.

You can also stratify your data. For example, if you are primarily interested in comparing three treatments, but suspect that there is a sex difference, you can divide the data into two strata, males and females. The final comparison of the treatments is a combined comparison of males with males and females with females. All the data for one stratum must be in the same table but different strata can be in different tables.

Results

Optionally, SURVIVCOMPARE displays a Kaplan-Meier plot (see the Public Procedure SURVIVGRAPH) of the survival curves for each group (with each stratum in a separate graph). It then returns a summary table of the number at risk, the observed number of events, the expected number of events, and the ratio of observed to expected for each group. If you are performing the Gehan-Breslow test the W-scores are included also. If the data are stratified, the above items are subdivided for each strata. In addition, the procedure shows the final chi-square statistic and significance level and optionally lets you see the variance-covariance matrix used to compute the test statistic.

Using SURVIVCOMPARE

To use the procedure type:

```
RAC PUBLIC SURVIVCOMPARE <GO>
CALL SURVIVCOMPARE <GO>
```

First it asks whether you would like to use the Mantel-Cox or Gehan-Breslow test and whether your data are stratified.

Unstratified Data

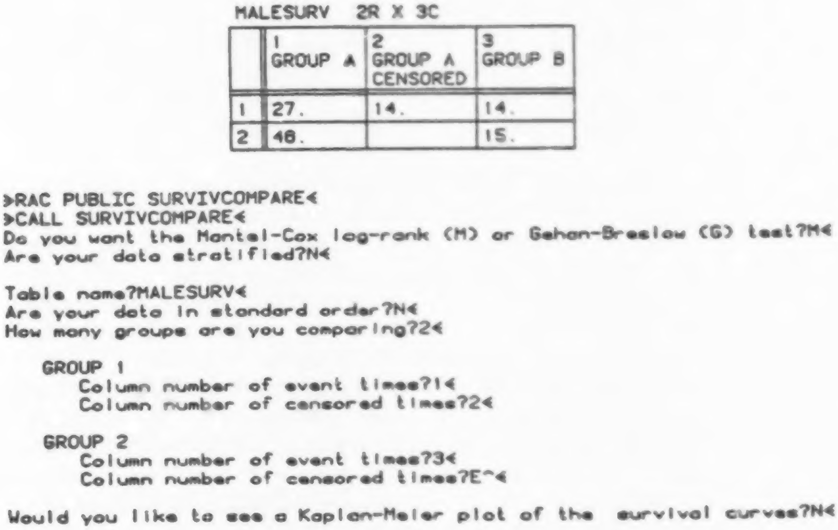
If your data are not stratified, you then are asked for the name of the data table and whether the data are in "standard" order.

Your data are in standard order if you want to use every column of your data table and columns are ordered so that:

column 1 contains group 1 event times,
column 2 contains group 1 censored times,
column 3 contains group 2 event times,
column 4 contains group 2 censored times,
...

If your data are not in standard order SURVIVCOMPARE asks explicitly for the columns of event and censored times for each group. Note that the data in Figure 1 are not in standard order because there is no column of censored data for group B.

Next the procedure asks whether you would like to see a Kaplan-Meier plot of the survival curves.



MALESURV 2R X 3C

	1 GROUP A	2 GROUP A CENSORED	3 GROUP B
1	27.	14.	14.
2	48.		15.

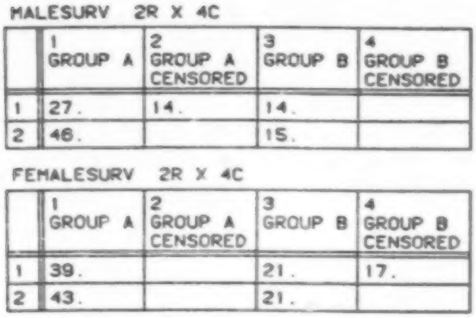
```
>RAC PUBLIC SURVIVCOMPARE<
>CALL SURVIVCOMPARE<
Do you want the Mantel-Cox log-rank (M) or Gehan-Breslow (G) test?M<
Are your data stratified?N<

Table name?MALESURV<
Are your data in standard order?N<
How many groups are you comparing?2<

   GROUP 1
      Column number of event times?1<
      Column number of censored times?2<

   GROUP 2
      Column number of event times?3<
      Column number of censored times?E^<

Would you like to see a Kaplan-Meier plot of the  survival curves?N<
```

Figure 1. SURVIVCOMPARE dialogue for a table not in standard order.

Stratified Data

If your data are stratified, SURVIVCOMPARE asks how many strata you have and whether all the strata are in the same table. Next it asks you for the name of the data table and whether the data are in standard order.

Your data can be in standard order in two ways depending on whether your data are in one or more tables. One way is for each stratum to be in a separate table in standard order (see Figure 2). The other way is for the strata to follow one another in the same table so that the first 2k columns contains the first stratum in standard order, the next 2k columns contains the second stratum in standard order, ..., where k is the number of groups (see Figure 3).



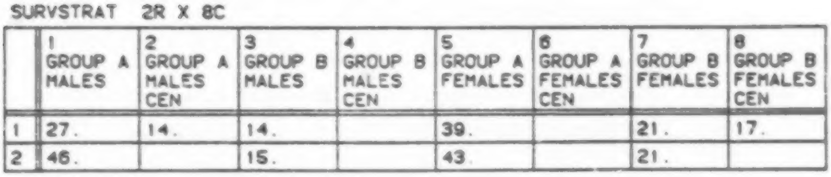
MALESURV 2R X 4C

	1 GROUP A	2 GROUP A CENSORED	3 GROUP B	4 GROUP B CENSORED
1	27.	14.	14.	
2	46.		15.	

FEMALESURV 2R X 4C

	1 GROUP A	2 GROUP A CENSORED	3 GROUP B	4 GROUP B CENSORED
1	39.		21.	17.
2	43.		21.	

Figure 2. Strata in separate tables in standard order.



SURVSTRAT 2R X 8C

	1 GROUP A MALES	2 GROUP A MALES CEN	3 GROUP B MALES	4 GROUP B MALES CEN	5 GROUP A FEMALES	6 GROUP A FEMALES CEN	7 GROUP B FEMALES	8 GROUP B FEMALES CEN
1	27.	14.	14.		39.		21.	17.
2	46.		15.		43.		21.	

Figure 3. Strata in one table in standard order.

For each stratum, SURVIVCOMPARE asks for the name of the data table (if in separate tables), the columns for event and censored times (if not in standard order), and whether you would like to see a Kaplan-Meier plot (see Figure 4).

```
>CALL SURVIVCOMPARE<
Do you want the Mantel-Cox log-rank (M) or Gehan-Breslow (G) test?G<
Are your data stratified?Y<
    How many strata do you have?2<
    Are the data for all the strata in the same table?N<

Table name for stratum 1?MALESURV<
Are your data in standard order?Y<

Would you like to see a Kaplan-Meier plot of the MALESURV survival cu-
rves?N<

Table name for stratum 2?FEMALESURV<

Would you like to see a Kaplan-Meier plot of the FEMALESURV survival -
curves?N<
```

Figure 4. Dialogue for stratified data in standard order.

Final Results Finally SURVIVCOMPARE displays the summary table and test results, and asks whether you would like to see the covariance matrix (see Figure 5).

```
Gehan-Breslow Test

MALESURV  7R X 5C  (NOTES)
```

0 GROUP	1 # AT RISK	2 # EVENTS OBSERVED	3 # EVENTS EXPECTED	4 OBSERVED/ EXPECTED	5 W SCORE
1. GROUP A	5.	4.	6.266667	.6382979	-9.
2. MALESURV	3.	2.	3.266667	.6122449	-5.
3. FEMALESURV	2.	2.	3.	.6666667	-4.
4.					
5. GROUP B	5.	4.	1.733333	2.307692	9.
6. MALESURV	2.	2.	.7333333	2.727273	5.
7. FEMALESURV	3.	2.	1.	2.	4.

```
Null Hypothesis H0:  All the groups have the same survival curve.

Chi-square = 6.08  D.F. = 1  P-value = .014

INFERENCE: REJECT NULL HYPOTHESIS AT P<0.05.


Would you like to see the covariance matrix?Y<

Variance-Covariance Matrix for Gehan-Breslow Test

MALESURV  2R X 2C
```

0 GROUP	1 GROUP A	2 GROUP B
1. GROUP A	13.33333	-13.33333
2. GROUP B	-13.33333	13.33333

Figure 5. Results of Gehan-Breslow test on stratified data in Figure 2.

Quiet entry
SURVIVCOMPAREQT The quiet entry to SURVIVCOMPARE returns the summary results table, covariance table, and the test statistic data. The calling sequence.is:

```
RAC PUBLIC SURVIVCOMPARE <GO>
CALL SURVIVCOMPAREQT (data-table, event-cols,
          censored-cols, mgtest, resultstab,
          covartab, df, chisq, siglev)
```

Argument	Type	Meaning/Value
data-tables	TABLE or TEXT array	Data table if one table or list of table names.
event-cols	FIXED array or FIXED array array	List of columns of event times. If stratified then Ith "row" is list for Ith stratum.
censored-cols	FIXED array or FIXED array array	List of columns of censored times. If stratified then Ith "row" is list for Ith stratum.
mgtest	TEXT	Either 'MANTEL-COX' ('M') or 'GEHAN-BRESLOW' ('G')
results-tab	TABLE	Table of summary results with test statistics in the notes.
covar-tab	TABLE	Variance-covariance table.
df	FIXED	Degrees of freedom.
chisq	FLOAT	Chi-square statistic.
siglev	FLOAT	Significance level.

If your data are not stratified then

- "data-tables" should be a TABLE,
- "event-cols" should be a FIXED array, and
- "censored-cols" should be a FIXED array.

The column numbers for the Jth group should be in the Jth elements of "event-cols" and "censored-cols". For example, for the data shown in Figure 1 you should use MALESURV, [1,3], and [2,].

If your data are stratified then

- "data-tables" should be either a TABLE or TEXT array,
- "event-cols" should be a FIXED array array, and
- "censored-cols" should be a FIXED array array.

When there is more than one table then the table name for the Ith stratum should be the Ith element of the TEXT array "data-tables". Event and censored column numbers for the Ith stratum should be in the Ith "row" of "event-cols" and "censored-cols", respectfully. The Jth group for all the strata should be in the Jth "column" of "event-cols" and "censored-cols". For example, for the data in Figure 3 you would use:

"data-tables" - SURVSTRAT
"event-cols" - [[1,3], [5,7]]
"censored-cols" - [[2,4], [6,8]]

NOTE: You should display "results-tab" with the NOTES qualifier.

References

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50:163-170, 1966.

Breslow N. A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. Biometrika 57:579-594, 1970.

2. SURVIVGRAPH

by
Karen Seo
Bolt Beranek and Newman Inc.

Summary:

SURVIVGRAPH provides an estimate of the probability of an event occurring, "mortality", or not occurring, "survival", using the product-limit (Kaplan-Meier) estimate. It produces a table of the cumulative distribution and a graph. More than one group at a time can be handled. A separate curve will be calculated for each group and added to the graph.

Calling Sequence:

```
RAC PUBLIC SURVIVGRAPH <GO>
CALL  SURVIVGRAPH <GO>
```

Entries:

```
CALL  SURVIVGRAPHQT (data,     event-cols,     censored-cols,
                     <table>   <fixed> or       <fixed> or
                             <fixed array>    <fixed array>

                     positive-slope?,      save?,
                        <bool>             <bool>

                     tables-for-distributions, graph)
                            <table> or         <graph>
                         <array of tables>
```

Citation:

No citation necessary.

Questions or Comments:

Call the PROPHET Hotline.

Things to remember:

1. The data (times of events) can be of type FLOAT, FIXED or NUMBER.
2. The graph and the tables containing the probability distribution are TEMPORARY.
3. There is also a public procedure SURVIVCOMPARE which compares survival curves.

best copy available

Background

"Survival" is a term used here to mean the time from a first well-defined event to a second event. The "survival curve" shows for each time t, the probability that the event of interest has NOT happened. Such a curve begins at a value of 1 at time 0, and decreases with time (i.e., has a negative slope) Similarly, a "mortality curve" shows for each time t, the probability that the event of interest HAS happened. Such a curve begins at a value of 0 at time 0, and increases with time (i.e., has a positive slope). The"product-limit" or "Kaplan-Meier" estimate is based on the notion that the probability of an event NOT occurring ("survival") until time t is equal to the probability of survival until time t-delta multiplied by the probability of surviving time t given that one has survived until time t-delta. Correspondingly, the probability of mortality is 1 minus the probability of survival. The required data consists of:

- time until event of interest for each subject
- "censored" times, i.e., time until "loss to follow-up" for any subject who somehow "left" the study before the event of interest occurred.

(For further statistical references, see:

1. Peto, et. al., Brit. J. Ca. V.34, p. 585 (1976) and V.35, p. 1 (1977)
2. BMDP-79 Manual, Chapter 19)

Description

SURVIVGRAPH calculates the Kaplan-Meier function for any number of groups (although, for more than 4 or 5, the graph can become cluttered). The data for all the groups must be in 1 table. For each group, the procedure takes 2 columns of the table (one for event times and one for censored times). It then performs the appropriate calculations and adds a probability function curve to the graph. A table of the function values can be displayed and placed in a temporary table . Also, a top-level copy of the graph can be obtained.

Data Set-up

The data must be set up in the following manner:

1. The data for all groups to be analyzed in a single execution of this procedure MUST be in 1 table.
2. For each group, there should be:
    + a column of event times of type FIXED, FLOAT or NUMBER
    + a column of censored times of type FIXED, FLOAT or NUMBER

   The column of censored times is optional. (So is the column of event times, but then you would not get any probability curve)

The columns for each group may be arranged in either of 2 ways:

1. The columns may be in any order within the table. In this case, you have to specify to the program the column numbers for each group.
2. The columns may be arranged in "standard" order. This means:
    + a column of event times for group 1 then
      a column of censored times for group 1 then
      a column of event times for group 2 then
      a column of censored times for group 2
      .
      .
      .
    + NO other kinds of columns.

    ALL columns must be present even if empty (eg. if there are no censored times for group 2, there still must be a censored column for group 2).

    In this case, you avoid having to specify to the program the columns for each group.

For an example of a data table see Figure 1. This table is set up in standard format.

Using SURVIVGRAPH

To call SURVIVGRAPH, give the command:

```
RAC PUBLIC SURVIVGRAPH <GO>
CALL  SURVIVGRAPH <GO>
```

The program will then ask you:

```
Name of table with event and censored data?
Do you want to use the entire table in standard format?
```

If you say NO, then it will prompt you:

```
How many groups are there?

Group1:
        Column number of column containing event times?
        Column number of column containing censored times?

Group2:
                Etc.
```

Then, whether you specified standard format or entered the columns for each group, it will ask:

```
Should the curve(s) represent:
     1. Probability that event has occurred ("Mortality")
     2. Probability that event has NOT occurred ("Survival")
Please enter "1" or "2" ?
```

Then, for each group, it will ask:

Do you wish to see the Kaplan-Meier function for group X?
Do you want a temporary table of this function for group X?

If you say YES, then it prompts:

Name for temporary table?

After calculating the Kaplan-Meier function for each group, it displays a graph containing a curve for each group. It then offers the option:

Do you want a top-level copy of this graph?

NOTE: The graph and the tables containing the function values for each group, are TEMPORARY. To keep them, you have to:
SAVE the graph
MAKE permanent tables from the temporary ones

Output

The procedure produces:

1. A graph with a separate probability curve for each subject group. A top-level copy may be obtained, which can then be edited and saved. The title of the graph will include the name of the data table and the type of function (survival or mortality). The curve labels will be taken from the column headings of the event columns. (See figure 2)
2. A table of the Kaplan-Meier function for each subject group. For each group, a separate temporary table containing the probability distribution may be obtained. It will have 4 columns (see figure 3):

Column 1:	Event time
Column 2:	Whether censored (C) or not (empty)
Column 3:	The probability value at that time
Column 4:	The standard deviation

Quiet Entry

This procedure can be called via its quiet entry:

1. No interactive input is necessary; all information is passed in as arguments. This makes this option well suited to use within a program.
2. It returns a graph and optionally, it returns tables containing the Kaplan-Meier function (one per group).

The arguments are:

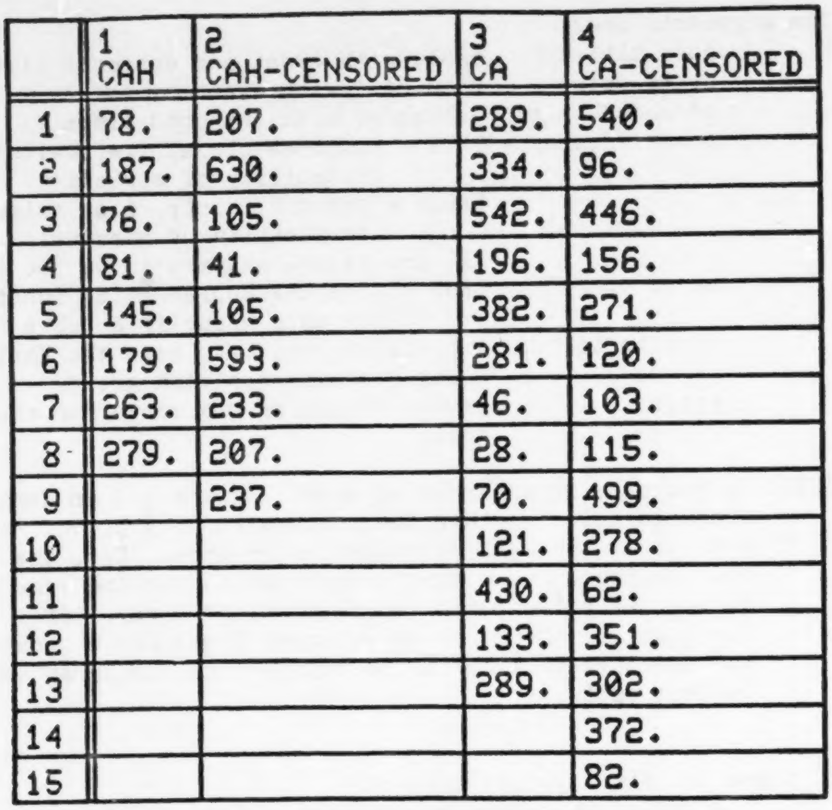
data table:	table with event and censored times
event cols:	column(s) with event times
censored cols:	column(s) with censored times
positive slope:	TRUE = positive slope, i.e. calculate probability of mortality FALSE = negative slope, i.e. calculate probability of survival All groups are assumed to be the same.
save:	TRUE = save the Kaplan-Meier function for each group in a table
results table:	table(s) in which to save the Kaplan-Meier function for each group
graph:	graph containing one curve for each group

NOTE: If you pass in an array of event columns and an array of censored columns, then the columns for any given group must be in the SAME position in the array. Eg., the 3rd col# in the event column array must be for the same group as the 3rd col# in the censored column array. If, for example, there are no censored times for a group, then the appropriate item in the censored column array would be left empty.

If there is only 1 group, then :

- the event column can be either a fixed number or an array of one item (i.e., 2 or 2])
- the censored column can be either a fixed number or an array of one item (i.e., 3 or 3])
- If the "save?" flag is TRUE, then the table in which to save the Kaplan-Meier function can be either a table or a table array.

Example of a Data Table



	1 CAH	2 CAH-CENSORED	3 CA	4 CA-CENSORED
1	78.	207.	289.	540.
2	187.	630.	334.	96.
3	76.	105.	542.	446.
4	81.	41.	196.	156.
5	145.	105.	382.	271.
6	179.	593.	281.	120.
7	263.	233.	46.	103.
8	279.	207.	28.	115.
9		237.	70.	499.
10			121.	278.
11			430.	62.
12			133.	351.
13			289.	302.
14				372.
15				82.

Figure 1

Example of a Graph

Do you want a top-level copy of this graph ?

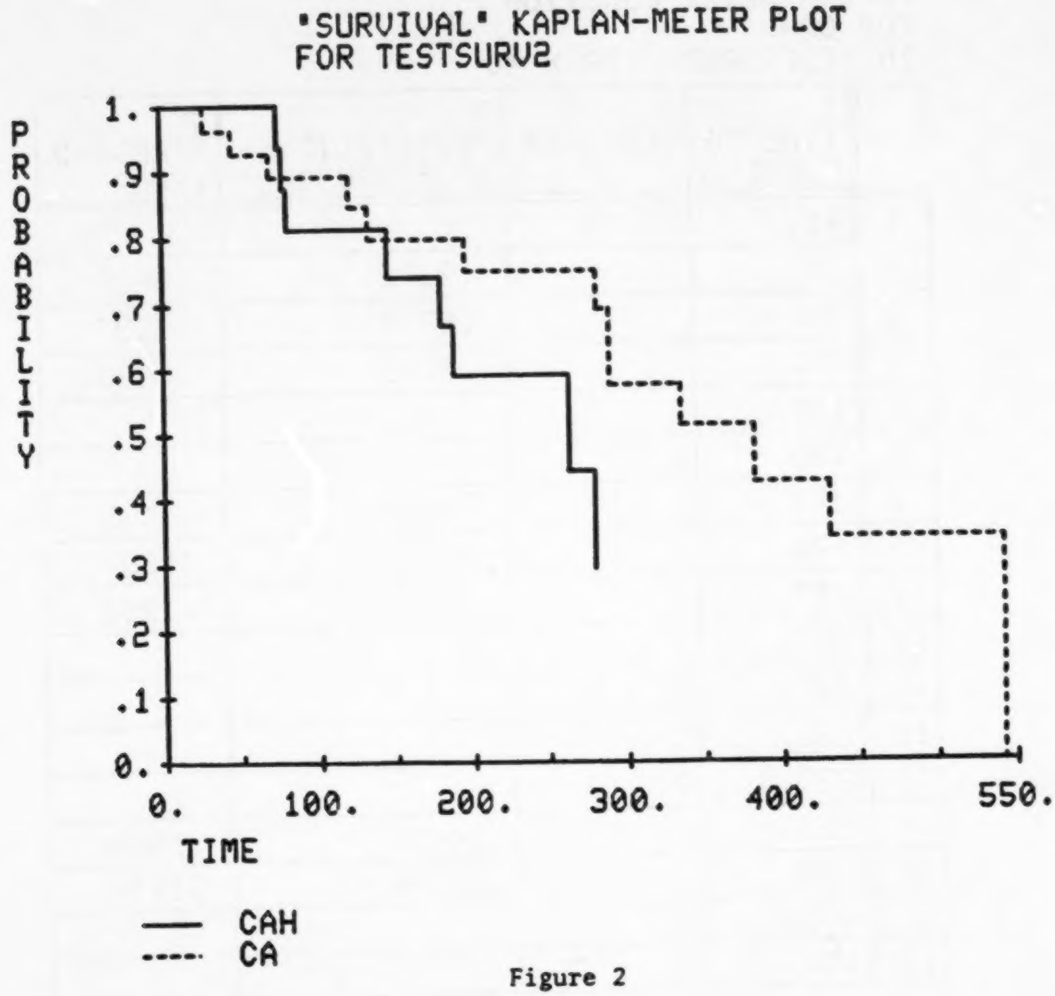


Figure 2

Example of Table Containing Kaplan-Meier Function

KAPLAN-MEIER FUNCTION
FOR CAH
IN TESTSURV2 17R X 4C

	1 TIME TO EVENT	2 CENSORED	3 PROBABILITY OF "SURVIVAL"	4 STANDARD ERROR
1	41.	C		
2	76.		.9375	.0605
3	78.		.8750	.0827
4	81.		.8125	.0976
5	105.	C		
6	105.	C		
7	145.		.7386	.1133
8	179.		.6648	.1237
9	187.		.5909	.1302
10	207.	C		
11	207.	C		
12	233.	C		
13	237.	C		
14	263.		.4432	.1609
15	279.		.2955	.1614
16	593.	C		
17	630.	C		

Figure 3

3. P2BMDP: Prophet-to-BMDP Interface

The public procedure P2BMDP is a PROPHET interface to the BMDP biostatistical package available on the PROPHET system. It allows you to easily convert tables to a form that the BMDP programs can use, add the appropriate BMDP control statements, and run specific BMDP programs without having to learn the PDP-10 Monitor commands. Using P2BMDP you can create a data file and a skeleton control file that appropriately describes this data to BMDP, change the control file to add statements required by the particular BMDP program you want to run, run BMDP and optionally display the output of results or obtain a lineprinter copy, and save or delete the files as you wish.

To access P2BMDP, enter the following commands PROPHET:

```
RAC PUBLIC P2BMDP  <GO>
CALL P2BMDP  <GO>
```

When P2BMDP starts, it prints the following:

```
Do you want to
1) Make a new control/data file set?             (M)
2) Change an old control file?                   (C)
3) Run a BMDP program?                           (R)
4) Delete a set of files?                        (D)
5) Scan a list of your BMDP files?               (S)
6) Exit this program?                            (E)
7) Type this text?                               (?)

M,C,R,D,S,E, or ?
```

Each function executes independently, so you can make several input files at once, use any of them as input to a BMDP program, delete those you don't want, or keep them for another PROPHET session. Typically you will want to make a control/data file set, change the control file to add additional control commands, run a BMDP program, and finally delete the control/data file set. Each function is described in more detail below.

1. M - If you type M P2BMDP makes a pair of BMDP files from a PROPHET table. PROPHET first asks for the name of the source table. You can exclude any rows or columns that you wish. Nonnumeric columns in your table are ignored, so you do not need to explicitly exclude them. PROPHET then asks for a name for the files. The name must be one to six alphanumeric characters. If a file already exists with that name, you will be asked for another.

   P2BMDP creates two files, a control file that contains the control statements expected by BMDP and a separate data file that contains the actual numbers from the table. The control file has a monitor name with an extension of ".BMD" and has a skeleton set of BMDP control statements. P2BMDP uses the table title in the BMDP problem title statement if the table has a title. Otherwise it uses the table name. For the variable names, P2BMDP uses the first eight characters in the column headings. It makes the names unique by numbering them when necessary. If the table has rownames, you can use them to identify individual cases for BMDP programs that use them. If the rownames are present, PROPHET asks you if you want to use zero, the first four, or the first eight characters for identifiers. The identifiers are ignored by BMDP programs which do not need them.

The second file has a filename identical with the control file, but has an extension of ".DAT" to distinguish it. It contains the actual data in the table, in a format suitably described by the control file. If your data contains empties, P2BMDP enters them as blanks in the data file and adds "BLANKS ARE MISSING" to the control file.

2. C - After you have created a set of input files, you will usually need to edit the control file to add any required or optional BMDP control statements needed for the BMDP program. When you are asked, type in the one-to-six character identifier for the control file you wish to edit. P2BMDP then puts you in the PROPHET editor described in the User's manual. If you wish, you can use the same set of data over and over, merely changing the appropriate commands in the command file for each different BMDP program.

3. R - To run a BMDP program type R. P2BMDP asks for the name of the control/data files you will use, and P2BMDP checks that they exist. After you specify the program to run P2BMDP starts it. Note that you must give the full program name (e.g., BMDP2V). When the BMDP program prints an asterisk prompt "*" you should type the name of the command file. (This is the only time you need to use the .BMD extention.) P2BMDP tells you what to type before PROPHET hands control to the program. When the BMDP program finishes, please note that YOU ARE NOT IN PROPHET, YOU ARE IN THE MONITOR.

To return to PROPHET and P2BMDP, you need only to type:

```
CALL RUNRTN <RETURN>
```

(P2BMDP reminds you to type this before it starts the BMDP program.)

When you return to P2BMDP, it asks you whether you would like the results printed on your terminal. Unfortunately, for most BMDP programs the output is formatted to fit on lineprinter paper and is too wide for the terminal screen. Therefore the output is difficult to read. The more recently developed BMDP programs are able to make the output 72 characters wide, as specified by the "/PRINT LINESIZE = 72." statement P2BMDP adds to the BMDP control file. In either case, you may find it useful to scan the output to check whether the BMDP program ran without errors. P2BMDP also asks whether you would like the output printed on the lineprinter and the listing mailed to you.

If your data or analysis require more space than provided by the BMDP programs, BMDP writes a message in the output file that it was "UNABLE TO ALLOCATE SUFFICIENT SPACE FOR ...". You should rerun the BMDP program using the R command. However, when you give the control file name in response to the BMDP "*" prompt you also need to include the /SPACE option. For example, suppose your command/date file set is called TEST. Then, when P2BMDP tells you to type:

```
TEST.BMD <RETURN>
```

you should type:

```
TEST.BMD/SPACE:10K <RETURN>
```

You can change the 10 to 15, 20, or any other number large enough so that your analysis has enough space. 10 will usually be sufficiant.

4. D - P2BMDP does not automatically delete your BMDP input file in case you want to run the same data again. You must use the Delete command to remove those you no longer need. Since the files may take up a large amount of space, it is recommended that you delete any unnecessary files. Please note that since they are closely related, P2BMDP deletes both the command file and data file.

5. S - You can type S to scan the list of the BMDP input files that you have created.

6. E - To exit P2BMDP and return to top-level PROPHET type E. You should be aware that your BMDP files remain unless you explicitly delete them using the D command. In this way you can reuse the same BMDP files from one PROPHET session to the next. You CANNOT archive or save your BMDP input files; they are not accessible from normal PROPHET.

Appendix A Update: Statistical Functions and Procedures

1. New quiet entries for Analysis of Variance, Kruskal-Wallis and Friedman's tests.

There are three new quiet entries which allow you to specify which rows and/or columns to exclude from the analysis. They are:

```
%ANOVASPEC
%KRUSKALWALLISSPEC
%FRIEDMANSPEC
```

They are similar to the previous quiet entries for these procedures, except that they have two additional arguments for specifying rows and columns.

The forms of the calls are:

```
CALL %ANOVASPEC(data, row-col, control-index,
                all-stats, non-sigs,
                block-spec, x-rows, x-cols)  <GO>
```

where:

data	is the TABLE containing the data
row-col	is a TEXT containing 'C' or 'R' (defaults to 'C')
control-index	is a FIXED specifying the control group, or empty for all-pairwise case
all-stats	is a BOOL specifying a full results table (TRUE) or an abbreviated results table (FALSE)
non-sigs	is a BOOL specifying display of all comparisons (TRUE) or only significant comparisons (FALSE)
block-spec	is a TEXT containing 'B' for blocked designs, empty otherwise

x-rows	is a FIXED ARRAY containing the rows to exclude
x-cols	is a FIXED ARRAY containing the cols to exclude

Note that when "x-rows" is empty, no rows will be excluded and similarly, when "x-cols" is empty, no columns will be excluded.

The arguments for the other two procedures are similar:

```
CALL %KRUSKALWALLISSPEC(data,row-col,control-index,
                all-stats,non-sigs,x-rows,x-cols)  <GO>

CALL %FRIEDMANSPEC(data,row-col,control-index,all-stats,
                non-sigs,x-rows,x-cols)  <GO>
```

Note that the old quiet entries, %ANOVA, %KRUSKALWALLIS, and %FRIEDMAN still exist.

All six of these quiet entries can be called so that PROPHET returns their results in a TABLE ARRAY. The first table in the array contains the results table from the overall test (e.g. analysis of variance table), and the second table in the array contains the results table from the multiple comparison test. For example,

```
tab-array = %ANOVA(data, row-col,
                   control-index, all-stats,
                   non-sigs, block-spec)  <GO>
```

Note that when you display the second table in the array (the multiple comparisons table) you will usually want to use the ROWHEIGHT=2 option as well as the NOTES option.

2. Quiet entries %JONCKHEEREQT and %PAGEQT

There are two quiet entries to %COMPAREORDERED that allow you to perform Jonckheere's and Page's tests within procedures. The calling sequences are:

```
res-tab = %JONCKHEEREQT (data, row-col,
              group-order, x-rows, x-cols)  <GO>

    res-tab = %PAGEQT (data, row-col,
              group-order, x-rows, x-cols)  <GO>
```

where:

res-tab	is the returned TABLE of results of the test
data	is the TABLE containing the data
row-col	is a TEXT containing 'C' or 'R' if EMPTY assumes 'C')
group-order	is a FIXED ARRAY containing the row or column numbers of the
groups	in order (if EMPTY then assumes same order as in table)
x-rows	is a FIXED ARRAY containing the rows to exclude (if EMPTY no rows are excluded)
y-rows	is a FIXED ARRAY containing the columns to exclude (if EMPTY no columns are excluded)

The results table should be displayed with the added options:

```
NOTES NORCSIZE NOCOLUMNNUMBERS
```

3. New results from %FITFUNQT and FITFUNQT

When you use the procedures %FITFUNQT and FITFUNQT after MLAB has completed a fit, the number of iterations the curve-fitter required will be reported, along with a statement on whether it converged. Also, if you enforced any constraints, a constraints table will be displayed which contains the associated Lagrange multipliers.

i. Quiet system procedure %FITFUNQT

```
F = %FITFUNQT(data-tab, x-col, y-col, wt-col
     func-to-fit,par-ivs, par-consts, fit-results,
     tab-opt, tab-results)  <GO>
```

where the first seven arguments are described in this manual, Part 3.

Fit-results should be declared a FLOAT array prior to the call. Upon return, it will contain the following:

[1] number of data points
[2] multiple R
[3] multiple R-squared
[4] standard error of regression
[5] number of iterations used by MLAB
[6] 1 if MLAB converged; 0 otherwise

Tab-opt is a FIXED which determines which fit table results are returned in tab-results.

Tab-results should be declared a TABLE array prior to the call. Upon return, if tab-opt=1, then tab-results will contain the following:

[1] parameter results table
[2] constraints table (only if there are constraints)
[3] analysis of variance table

If tab-opt=2 then tab-results will contain the following:

[4] table of residuals

If tab-opt=3, then elements [1], [2], [3] and [4] of tab-results will be set as above. If tab-opt has any other value, NO table results will be returned.

ii. Quiet entry to the Public Procedure FITFUN

```
F = %FITFUNQT(function-tab, pars-tab, data-tab,
    ind-var-cols, par-vary, par-consts, fit-results,
    tab-opt, tab-results) <GO>
```

The first six arguments are described in the PROPHET STATISTICS Manual, part 3, and fit-results should be declared a FLOAT array array prior to the call. Upon return, it will contain the following in the ith row (corresponding to the ith function in the system):

[I,1] number of data points
[I,2] multiple R
[I,3] multiple R-squared
[I,4] standard error of regression
[I,5] number of iterations used by MLAB
[I,6] 1 if MLAB converged; 0 otherwise

Tab-opt has the same meaning as above (see %FITFUNQT).

Tab-results should be declared a TABLE array array prior to the call. Upon return, it will contain the following in the ith row (corresponding to the ith function in the system):

If tab-opt=1 THEN

[I,1] parameter results table
[I,2] constraints table (only if there are constraints)
[I,3] analysis of variance table

If tab-opt=2 then the ith row of tab-results will contain the following:

[I,4] table of residuals

If tab-opt=3, then elements [I,1], [I,2], [I,3] and [I,4] will be set as above. If tab-opt has any other value, NO table results will be returned.

4. %FITMULTIPLEQT

%FITMULTIPLEQT is the "quiet" form of the FIT MULTIPLE command. You call the procedure as follows:

```
%FITMULTIPLEQT(data-tab, y-col, x-cols, fit-stats,
               tab-option, tab-results) <GO>
```

%FITMULTIPLEQT performs a QUIET fit, i.e., once called, it does not ask any questions nor display any results. This procedure performs a single fit using all the specified X terms. You can obtain any fit results you desire. Notice that the arguments are very similar to those of the other QUIET fitting system procedures:

data-tab	TABLE (input) contains observed data (must have NO empty cells)
y-col	FIXED (input) column of Y data
x-cols	FIXED array (input) columns of X data
fit-stats	FLOAT (output) after the fit is completed, this array contains:

[1] number of data points
[2] multiple R
[3] multiple R-squared
[4] standard error of regression
[5] Y intercept

Tab-option FIXED (input) determines what table results to return in tab-results. If tab-option = 1, then elements 1 and 2 equal the parameter and ANOVA tables. If tab-option=2, then element 3 equals the table of residuals. If tab-option=3 then elements 1, 2 and 3 are the parameter, ANOVA, and residuals tables. If tab-option equals any other value, then no table results are returned.

5. BLISSPROBIT

by

Mark B. Sostrin

(with statistical guidance by Willard J. Lennox)
United States Army Medical Research Institute
of Chemical Defense

Summary

BLISSPROBIT uses the Bliss method of maximum likelihood to compute the probit equation. The procedure offers two modes. The first involves the computation of probits and associated statistics for a single data set of quantal dose-response values. Included are the observed probits, expected probits, the percent response relating to the expected probits, the doses producing the ED01, ED16, ED30, ED50, ED84, and ED99 percent response values with their respective 95% confidence limits plus graphs of observed and expected probit vs log dose and the percent response vs log dose.

The second mode performs parallel assays by first computing probits (as above) for a standard data set of dose-response values plus related data sets. Next, appropriate statistical tests for linearity and parallelism are performed and the common slope is derived and new probit equations are computed based upon the common slope. The new ED50 levels and their 95% confidence limits are then computed for the standard and each unknown. Finally the relative potency is computed with its 95% confidence limits and the reciprocal of the relative potency is also shown for use in working with protective ratios.

Calling Sequence

```
RAC PUBLIC BLISSPROBIT  <GO>
   CALL BLISSPROBIT  <GO>
```

Citation

Sostrin, Mark B., BLISSPROBIT, in USAMRICD PROCEDURES Notebook, USAMRICD, APG, MD., 1983.

Questions or Comments
contact Mark B. Sostrin (USAMRICD) (301)-671-3184

After calling the procedure you are asked which mode you would like to use: whether probits for a single data set or parallel assays for a standard and unknown data sets. The procedure assumes that each data set is saved in a table such that

col 1 has the doses
col 2 has the number responding
col 3 has the number in the sample.

Use the "SET TITLE OF <TABLE>" command to give a descriptive label to each data set.

The following results are presented in the probit analysis of each data set:

i. observed probits
ii. expected probits
iii. percent response relating to the expected probits
iv. doses producing ED01, ED16, ED30, ED50, ED84, ED99 and their respective 95% confidence limits
v. graph of observed and expected probit vs log dose
vi. graph of percent response vs log dose.

If you are performing parallel assays, you are asked to specify the standard data set first and then the table names of all the related data sets. You are also asked to indicate the natural mortality ratio for each data set. The procedure continues by showing the above results for each data set. Next, appropriate statistical tests for linearity and parallelism are performed and the common slope is derived and new probit equations are computed based upon the common slope. The new ED50 levels and their 95% confidence limits are then computed for the standard and each unknown. Finally the relative potency is computed with its 95% confidence limits and the reciprocal of the relative potency is shown for use in working with protective ratios.

For convenience, you can obtain "top-level" copies of any table or graph result. Refer to the examples at the end for further illustration.

BLISSPROBIT Computations

BLISSPROBIT uses the Bliss [see 1.] method of maximum likelihood to calculate the probit equation. In the Bliss method of computing the dosage-effect curve, each observation is assigned a weight (based upon the variance of the probit response) and the empirical probits are replaced by working probits. The variance of the probit is a function of the response expected at that dose, so that

v(p)=PQ/N

p is the expected value of the observed p
Q = 1-P

Each probit is given a weight (w) equal to the reciprocal of its variance or

w = N*(Z**2)/(P*Q)

Z is the ordinate of the normal curve at the expected proportionate response P.

Each working probit (y) is computed from the equation:

y = (Y-P/Z)+p*(1/Z) = (Y+Q/Z)-q*(1/Z)

where p is the proportion which was observed to react and P and Z are the proportionate area and the ordinate of the normal curve corresponding to the provisional expected probit Y.

The corrected dosage-effect curve is then computed from the values of w,x, and y.

In the following equations, S( ) stands for the summation of the enclosed expression.

XBAR=S(w*x)/S(w) and YBAR=S(w*y)/S(w) weighted means

[wx2]=S(w*x**2)-XBAR*S(w*x) weighted sum of squares

[wxy]=S(w*x*y)-YBAR*S(w*x) product of deviations from the two means

b=[wxy]/[wx2] slope of dosage-effect curve

[wy2]=S(w*y**2)-YBAR*S(w*y) weighted total variation in the response

B2=[wxy]**2/[wx2] effect of the straight line

CHI-SQUARE=[wy2]-B2 chi-square for linearity

df= k-2 degrees of freedom,
where k is the number of doses.

All doses are included in the calculation of k regardless of the magnitude of the expected percent response for the particular dose. This may vary from Bliss's approach of ignoring those points which have an expected number in the smaller class, either positive or negative, less than 0.5 individuals.

Note that a significantly large chi-square causes all variances to be modified by a heterogeneity factor equal to the chi-square divided by its degrees of freedom. In all uses of the standard errors, the t-distribution with (k-2) df is used instead of the normal distribution.

SE(a)=SQRT(1/S(w)) standard error of position

SE(b)=SQRT(1/[wx2]) standard error of slope

Sx=LAMBDA*SQRT((1/S(w)+(Y-YBAR)**2/(B2-T**2)))
standard error of a log dose

XL=XBAR+LAMBDA*C**2*(Y-YBAR)+/-(T*C*Sx)
confidence limits for x at a given y

where C**2=B2/(B2-T**2)
and LAMBDA=1/b.

Expected proportions and normal equivalent deviates (NED) are calculated using algorithms expounded by Hastings [see 2.]. The inverse student's t-function, T, is formulated using the Cornish-Fisher [see 3.] asymptotic expansion.

Expected proportions are calculated by the formula:

PW=SQRT(2)*AMU*Z*(C1+AMU*(C2+AMU*(C3+AMU*(C4+C5*AMU))))

where AMU=1/(0.23164189*ABS(NED)+1)

and NED=EXPECTED PROBIT-5.0

and

C1= 0.22583685
C2=-0.25212870
C3= 1.25965910
C4=-1.28782240
C5= 0.94064607

The computation for computing the NED from an observed proportion is:

NED=R-(Co+C1*R+C2*R**2)/(1+D1*R+D2*R**2+D3*R**3)+DELTA(e)

where 0 <= e <= 0.5

e = Q = 1-P, where P = observed proportion

and R = SQRT((LOG(1/e**2))

C0=2.515517
C1=0.802853
C2=0.010328
D1=1.432788
D2=0.189269
D3=0.001308

The inverse student's t-function, T, is formulated by:

T=U+(U**3+U)/4*v+(5*U**5+16*U**3+3*U)/96*v**2+
(3*U**7+19*U**5+17*U**3-15*U)/384*v**3+
(79*U**9+776*U**7+1482*U**5-1920*U**3-945*U)/92160*v**4+
(27*U**11+339*U**9+930*U**7-1782*U**5-765*U**3+17955*U)/
368640*v**5+........

where v = df of t
U= standard normal variate
(in the program, U=1.96)

Computation tests whether the use of a combined assay slope is valid: The common slope is determined as follows:

Bcom=(S[wxy])/(S[wx2])

The total variation in y accounted for by Bcom is:

Bc2=(S[wxy])**2/S[wx2]

The assay chi-square is broken down into components for heterogeneity and parallelism.

Chi-square for linearity =S(chi-squares from individual lines)
Degrees of freedom = S(df from individual lines)

```
Chi-square for parallelism =S(B2)-Bc2
Degrees of freedom = number of lines - 1
```

The log relative potency is computed by:

```
M=Xs-Xu
```

where Xs and Xu are the LOG ED50's for the standard and the unknown, respectively.

Therefore, the relative potency is `EXP(2.3025851*M)`

The standard error of M can be computed as:

```
Sm=(SQRT(S**2)/Bcom)*SQRT(1/S(w)u+1/S(w)s+
    (YBARu-YBARs)**2/(Bc2-S**2*T**2))

where S**2=CHI-SQUARE/DF
```

Note that for non-heterogenous data, `S**2` is assumed to be 1.0.

The confidence limits are computed by the equation:

```
XL=XBARs-XBARu-(C**2*(YBARs-YBARu))/Bcom +/- T*C*Sm

where C**2=Bc2/(Bc2-S**2*T**2)
and C=SQRT(C**2)
```

References

1. Bliss, C.I. THE STATISTICS OF BIOASSAY with special reference to the vitamins, reprinted, with additions from VITAMIN METHODS VOL II, ACADEMIC PRESS INC., PUBLISHER, NEW YORK, 1952.

2. Hastings, C. jr. APPROXIMATIONS FOR DIGITAL COMPUTERS, PRINCETON UNIVERSITY PRESS, PRINCETON, NEW JERSEY, 1955.

3. Fisher, R.A., Cornish, E.A., TECHNOMETRICS, 2, 209, 1960.

4. Abramowitz, M., Stegun, I.A., HANDBOOK OF MATHEMATICAL FUNCTIONS WITH FORMULAS, GRAPHS, AND MATHEMATICAL TABLES, NBS APPLIED MATHEMATICS SERIES 55, 1964.

5. Finney, D.J., PROBIT ANALYSIS, 3rd EDITION, CAMBRIDGE UNIVERSITY PRESS, 1971.

PROPHET Statistics Manual:

INDEX

Page

ADD MEASURES 1-14
ADD STATS 1-9
%ADDFITTEDVALUES procedure B-15
Analysis of covariance 8-1
Analysis of variance 4-4, 4-6, 4-10, 4-13, 7-1, B-16
Analysis of variance in regression 6-6
ANOVA procedure B-16
%ANOVA procedure A-2
Ansari-Bradley test 3-4, A-5
ANVAR procedure 7-3
ARAFIL procedure 7-10, 7-12

Bartlett's test 4-25, A-32
Binomial distribution 2-13, A-6
Blocked design 4-3, 4-13
BMDP Interface with PROPHET . . . . B-45 - B-48

CDIST procedure B-3
Chi-squared distribution 2-13, A-3, A-6
Chi-squared test for goodness of fit 5-20
Chi-squared test for independence 5-2, 5-4, 5-11
%CHISIGLEV procedure A-3
CHISQ command 5-2, 5-20
Coefficient of variation 1-11
COMPARE 3-3, 4-2
%COMPAREORDERED procedure B-9
Comparing elevations 8-6, 8-12
Comparing regressions (see TEST LINES)
Comparing slopes 8-5, 8-11
CONFCURVS procedure 6-15
CONFCURVSQT procedure A-4
Confidence curves 6-11, 6-15, A-4
Confidence intervals 1-3, 1-19
Constraints in regression 6-27
Contingency coefficient 5-2, 5-5, 5-11
Contingency tables 5-1
Control group 4-4, 8-10
Convergence 6-26
Correlation coefficient 6-13
Correlation coefficient squared 6-9
%CORRELATION function B-11
CORRELATIONS expression B-11
Covariance analysis 8-1
COVARIANCES expression B-11
Cramer's V statistic 5-2, 5-11
CROSSTAB procedure 5-9, B-10
CV% statistic 1-11

D'Agostino's test 2-6, 2-9, A-24
DELETE STATS command 1-10

Page

%DISPTEST procedure . . . A-5
%DISTRIBUTION procedure . . . A-6
Distribution tables . . . 2-2
%DOKSUM . . . A-8
%DUNCAN procedure . . . B-6
%DUNNET procedure . . . B-6
Dunnett's test . . . 4-5, 4-16, 4-24, A-23

Empties in analysis of variance . . . 4-15, 7-5
EXPFIT procedure . . . 6-5, 6-35
Exponential regression . . . 6-35

F-distribution . . . 2-13, 3-10, 4-7, A-6, A-21
F-test for equality of variances . . . 3-16, A-22
Factorial analysis of variance . . . 7-1
Factors . . . 7-1
FDATA2A procedure . . . B-1
Fisher procedure . . . 5-7
Fisher's exact test . . . 5-7
FIT FUNCTION command . . . 6-5, 6-22
FIT FUNCTION procedure . . . B-13
FIT LINE command . . . 6-5, 6-13
FIT MULTIPLE command . . . B-11
FIT POLYNOMIAL command . . . 6-5, 6-23
FITFUN procedure . . . 6-5, 6-30, B-13
FITFUNQK procedure . . . 6-33
FITFUNQKSAV procedure . . . 6-34
%FITFUNQT procedure . . . A-9
FITFUNQT procedure . . . A-12
FITFUNSAV procedure . . . 6-33
%FITLINEQT procedure . . . A-16
%FITPOLQT procedure . . . A-18
Frequency Data and Descriptive Statistics . . . B-1
Frequency tables . . . 5-1
%FRIEDMAN procedure . . . A-20
Friedman's test . . . 4-5, 4-21, A-20
%FSIGLEV procedure . . . A-21
%FTEST . . . A-22
Full results table . . . 4-18, 8-12

Gamma statistic . . . 5-11, 5-18
Geary's statistic . . . 2-8
GEOMEAN statistic . . . 1-12
Geometric mean . . . 1-12
Graphs . . . 1-19

Histograms . . . 2-2, 2-5
Hotelling procedure . . . 9-4
Hotelling's test . . . 9-2, 9-4
HYPERBOLIC procedure . . . 6-5, 6-40
Hyperbolic regression . . . 6-40

Page
Independence 5-2, 5-4
Interactions 7-2

Kendall's tau statistic 5-11, 5-17
Kolmogorov-Smirnov test 2-11, A-6
Kruskal-Wallis test 4-5, 4-19, A-23
%KRUSKALWALLIS procedure A-23
Kurtosis 1-13, 2-5, 2-7, 2-9, A-25

Lambda statistic 5-11, 5-14
Levels 7-1
Levene's test 4-4, 4-25, A-32
Linear regression 6-6, A-16
Lineweaver-Buke plot 6-41
LOGMEASURES command 1-17
LOGSTATS command 1-17

Main effects 7-2
MAKE TABLE AS DISTRIBUTION command 2-2
%MAKEDISTABLE command B-3
Mann-Whitney test 3-14, A-27
Maximum 1-12, 1-18
Mean 1-2, 1-18
MEASURES command 1-14, 1-18, A-28
Median 1-11, 1-18
Michaelis-Menton equation 6-40
Mid-range 1-12
Minimum 1-12, 1-18
Missing values (see Empties)
MLAB 6-26, B-13
%MULTICOMP procedure B-6
Multiple comparisons 4-1, 4-5, 4-16, 4-23, 8-9, B-6
Multiple correlation coefficient squared 6-9
Multiple linear regression 6-16, B-11
Multivariate analysis 9-1

Newman-Keuls test 4-5, 4-16, 8-6, 8-7, A-2
Nominal data 5-1, 5-14
Nonparametric tests 2-1, 3-2
Normal distribution 2-5, 2-6, 2-9, 2-13, A-6
A-24, A-25, A-26
%NORMALITY procedure A-24
%NORMALSHAPE procedure A-25
%NORMSIGLEV procedure A-26

One-way analysis of variance 4-10, A-2
Ordered Ranks, for data in tables B-1
Ordinal scale data 3-2, 5-1, 5-14
Outliers 1-19, 6-10

P2BMDP procedure B-45
Parametric tests 2-1, 3-2
Peeling in exponential regression 6-35

Page

Poison distribution 2-13, A-6
Polynomial regression 6-22, A-18
PSCALE procedure B-3

Range 1-12
RANKS expression B-1
%RANKTEST procedure A-29
Regression 6-1, 6-4, 6-5
Regression through the origin 6-12
Residuals 6-7, 6-10

Scatter plot 6-2
%SCHEFFE procedure B-6
SELECT STATS command 1-15
SEM (standard error of the mean) 1-11
Skewness 1-12, 2-5, 2-7, 2-9, A-25
%SNK procedure B-6
Spearman's rho statistic 5-11, 5-16
Standard deviation 1-2, 1-18
Standard error of regression 6-9
STATPAK 5-11, 5-14
STATS command 1-4, A-28
%STATSDP variable A-28
%STATSLIST variable A-28
SURVIVCOMPARE procedure . . . . B-30 - B-36
SURVIVGRAPH procedure . . . . B-37 - B-44

t-distribution 2-13, 3-8, A-6, A-29, A-30
t-test,equal variance 3-10, A-31
t-test,one-sample 3-6, A-31
t-test,paired 3-9, A-31
t-test,unequal variance 3-11, A-31
TEST DISTRIBUTION command 2-13
TEST LINES command 8-9
TEST NORMALITY command 2-9
TEST NORMALSHAPE command 2-9
TEST VARIANCES command 3-16, 4-25
%TFUN procedure A-28
%TSIGLEV procedure A-29
%TTEST procedure A-30
%TUKEY procedure B-6
Two-way analysis of variance 4-13, A-2

Uniform distribution 2-13, A-6

Variance 1-2, 1-18
%VARIANCETEST procedure A-32

Weighted regression 6-12, 6-14, 6-35
Wilcoxon-Mann-Whitney test 3-14, A-27
Wilcoxon signed rank test, one-sample 3-13, A-27
Wilcoxon signed rant test, paired 3-14, A-27
Wilk-Shapiro test 2-6, 2-9, A-24